Evolution of a Cytoplasmic Determinant: Evidence for the Biochemical Basis of Functional Evolution of the Novel Germ Line Regulator Oskar

Leo Blondel [ID],[1] Savandara Besse [ID],[†,1] Emily L. Rivard [ID],[1] Guillem Ylla [ID],[‡,2] and Cassandra G. Extavour [ID][*,1,2]

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA
[2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA
[†]Present address: Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal, QC, Canada
[‡]Present address: Laboratory for Bioinformatics and Genome Biology, Jagiellonian University, Krakow, Poland
*Corresponding author: E-mail: extavour@oeb.harvard.edu.
Associate editor: Harmit Malik

Abstract

Germ line specification is essential in sexually reproducing organisms. Despite their critical role, the evolutionary history of the genes that specify animal germ cells is heterogeneous and dynamic. In many insects, the gene *oskar* is required for the specification of the germ line. However, the germ line role of *oskar* is thought to be a derived role resulting from co-option from an ancestral somatic role. To address how evolutionary changes in protein sequence could have led to changes in the function of Oskar protein that enabled it to regulate germ line specification, we searched for *oskar* orthologs in 1,565 publicly available insect genomic and transcriptomic data sets. The earliest-diverging lineage in which we identified an *oskar* ortholog was the order Zygentoma (silverfish and firebrats), suggesting that *oskar* originated before the origin of winged insects. We noted some order-specific trends in *oskar* sequence evolution, including whole gene duplications, clade-specific losses, and rapid divergence. An alignment of all known 379 Oskar sequences revealed new highly conserved residues as candidates that promote dimerization of the LOTUS domain. Moreover, we identified regions of the OSK domain with conserved predicted RNA binding potential. Furthermore, we show that despite a low overall amino acid conservation, the LOTUS domain shows higher conservation of predicted secondary structure than the OSK domain. Finally, we suggest new key amino acids in the LOTUS domain that may be involved in the previously reported Oskar−Vasa physical interaction that is required for its germ line role.

Key words: oskar, vasa, Drosophila, germ plasm, germ cell, LOTUS domain, RNA binding, hidden Markov models, Hymenoptera, Lepidoptera, Zygentoma.

Introduction

With the evolution of obligate multicellularity, many organisms faced a challenge considered a major evolutionary transition: allocating only some cells (germ line) to pass on their genetic material to the next generation, relegating the remainder (soma) to death upon death of the organism (reviewed in Kirk [2005]). Although there are multiple mechanisms of germ cell specification, they can be grouped into two broad categories, induction or inheritance (reviewed in Extavour and Akam [2003]). Under induction, cells respond to an external signal by adopting germ cell fate. Under the inheritance mechanism, maternally synthesized cytoplasmic molecules, located within a specialized cytoplasm called germ plasm, are deposited in the oocyte and "inherited" by a subset of cells during early embryonic divisions. Cells inheriting these molecules commit to a germ line fate (reviewed in Extavour and Akam [2003]).

The inheritance mechanism in insects that undergo metamorphosis (Holometabola) appears to have evolved by co-option of a key gene, *oskar*. *oskar* was first identified in forward genetic screens for axial patterning mutants in *Drosophila melanogaster* (Lehmann and Nüsslein-Volhard 1986). For the first 20 years following its discovery, *oskar* appeared to be restricted to Drosophilids (Clark et al. 2007). Its later discovery in the mosquitoes *Aedes aegypti*, *Anopheles gambiae*, and *Culex quinquefasciatus* (Juhn and James 2006; Juhn et al. 2008) and the wasp *Nasonia vitripennis* (Lynch et al. 2011) suggested the hypothesis that *oskar* emerged at the base of the Holometabola, and facilitated the evolution of germ plasm in these insects (Lynch et al. 2011). However, our subsequent identification of *oskar* homologs in the cricket *Gryllus bimaculatus* (Ewen-Campen et al. 2012), and in many additional hemimetabolous insect species (Blondel et al. 2020), demonstrated that *oskar* predates the

Open Access

Holometabola, and must be at least as old as the major radiation of insects (Misof et al. 2014). Two secondary losses of *oskar* from insect genomes have also been reported, in the beetle *Tribolium castaneum* (Lynch et al. 2011) and the honeybee *Apis mellifera* (Dearden et al. 2006), and neither of these insects appear to use germ plasm to establish their germ lines (Nelson 1915; Nagy et al. 1994; Dearden 2006; Schroder 2006). Whether *oskar* is ubiquitous across all insect orders, whether it is truly unique to insects, the evidence for or against potential losses or duplications of the *oskar* locus across insects, and the evolutionary dynamics of the locus, remain unknown.*oskar* remains, to our knowledge, the only gene that has been experimentally demonstrated to be both necessary and sufficient to induce the formation of functional primordial germ cells (called pole cells in *Drosophila*) (Kim-Ha et al. 1991; Ephrussi and Lehmann 1992). Thus, in *D. melanogaster* (Lehmann and Nüsslein-Volhard 1986; Kim-Ha et al. 1991; Ephrussi and Lehmann 1992) and potentially more broadly in holometabolous insects with germ plasm (Lynch et al. 2011; Rafiqi et al. 2020), *oskar* plays an essential germ line role. However, it is clear that *oskar*'s germ line function can evolve rapidly, as even within the genus *Drosophila*, *oskar* homologs from different species cannot always substitute for each other (Webster et al. 1994; Jones and Macdonald 2007). Moreover, the ancestral function of this gene may have been in the nervous system rather than the germ line (Ewen-Campen et al. 2012). The current hypothesis is therefore that it was co-opted to play a key role in the acquisition of an inheritance-based germ line specification mechanism ∼300 Mya (Misof et al. 2014), in the lineage leading to the Holometabola (Ewen-Campen et al. 2012). Thus, the case of *oskar* offers an opportunity to study the evolution of protein function at multiple levels of biological organization, from the genesis of a novel protein, through to potential co-option events and the evolution of functional variation.

Neofunctionalization often correlates with a change in the fitness landscape of the protein sequence caused by novel biochemical constraints imposed by amino acid sequence changes (Sikosek et al. 2012; Sikosek and Chan 2014). Such potential constraints may be revealed by analyzing the conservation of amino acids, their chemical properties, or structure at the secondary, tertiary or quaternary levels (Sikosek and Chan 2014). Oskar has two well-structured domains conserved across identified homologs to date (Blondel et al. 2020): an N-terminal Helix Turn Helix domain termed LOTUS with potential RNA-binding properties (Anantharaman et al. 2010; Jeske et al. 2015; Yang et al. 2015; Jeske et al. 2017), and a C-terminal GDSL-lipase-like domain called OSK (Jeske et al. 2015; Yang et al. 2015) (fig. 1). These two domains are linked by an unstructured highly variable interdomain sequence (Ahuja and Extavour 2014; Jeske et al. 2015; Yang et al. 2015). We previously showed that this domain structure is likely the result of a horizontal transfer event of a bacterial GDSL-lipase-like domain, followed by the fusion of this domain with a LOTUS domain in the host genome (Blondel et al. 2020). Biochemical assays of the properties of the LOTUS and OSK domains provide some clues as to the molecular mechanisms that Oskar uses

to assemble germ plasm in *D. melanogaster*. The LOTUS domain is capable of homodimerization (Jeske et al. 2015, 2017), and directly binds and enhances the helicase activity of the ATP-dependent DEAD box helicase Vasa, a germ plasm component (Jeske et al. 2017). The OSK domain resembles GDSL lipases in sequence (Jeske et al. 2015; Yang et al. 2015; Blondel et al. 2020), but is predicted to lack enzymatic activity, as the conserved amino acid triad (S200 D202 H205) that defines the active site of these lipases is not conserved in OSK (Anantharaman et al. 2010; Jeske et al. 2015; Yang et al. 2015). Instead, copurification experiments suggest that OSK has RNA-binding properties, consistent with its predicted basic surface residues (Jeske et al. 2015; Yang et al. 2015). Whether or how changes in the primary sequence of Oskar can explain the evolution of its molecular mechanism or tissue-specific function, remain unknown.

To date, sequences of ∼100 *oskar* homologs have been reported (Lynch et al. 2011; Jeske et al. 2015; Quan and Lynch 2016; Blondel et al. 2020). However, the vast majority of these are from the Holometabola, and it is thus unclear whether analysis of these sequences alone would have sufficient power to allow extrapolation of conservation and divergence of putative biochemical properties across insects broadly speaking. Multiple hypotheses as to the molecular mechanistic function of particular amino acids in the LOTUS and OSK domains in *D. melanogaster* have been proposed (Jeske et al. 2015; Yang et al. 2015; Jeske et al. 2017), but without sufficient taxon sampling, the potential relevance of these mechanisms to *oskar*'s evolution and function in other insects is unclear.

Here we address these outstanding questions by applying a rigorous bioinformatic pipeline to generate the most complete collection of *oskar* sequences to date. By analyzing 1,862 Pancrustacean genomes and transcriptomes, we show that *oskar* likely first arose at least 400 Ma, before the advent of winged insects (Pterygota). We find that the *oskar* locus has been lost independently in some insect orders, including near-total absence from the order Hemiptera, and clarify that the absence of *oskar* from the *Bombyx mori* and *T. castaneum* genomes (discussed in Quan and Lynch 2016) does not reflect a general absence of *oskar* from Lepidoptera or Coleoptera. By comparing Oskar sequences in a phylogenetic context, we reveal that distinct biophysical properties of Oskar are associated with Hemimetabola and Holometabola. We use these observations to propose testable hypotheses regarding the putative biochemical basis of evolutionary change in Oskar function across insects.

Results

HMM-Based Discovery Pipeline Yields Hundreds of Novel *oskar* Homologs

We wished to study the evolution of the *oskar* gene sequence as comprehensively as possible across all insects. To expand our previous collection of nearly 100 homologous sequences (Blondel et al. 2020), we designed a new bioinformatics pipeline to scan and search for *oskar* homologs across all 1,565
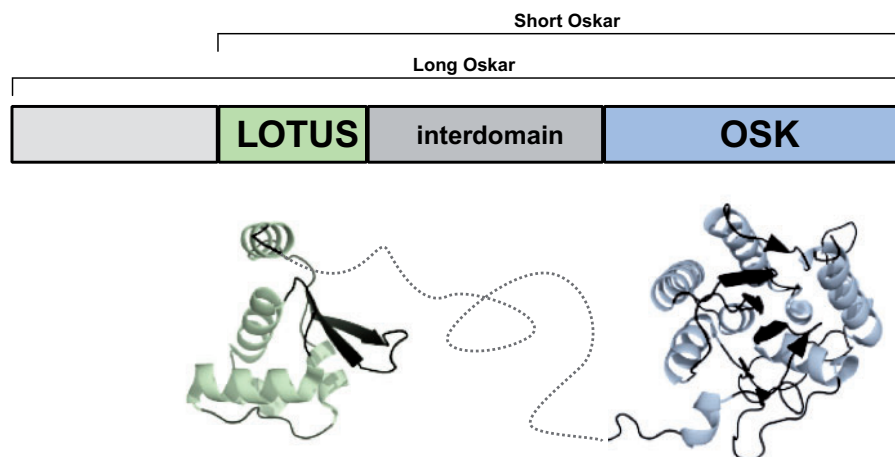
FIG. 1. Overview of Oskar protein structure. The most common isoform of the Oskar protein, Short Oskar, is composed of two well-folded domains, LOTUS and OSK, separated by an interdomain sequence. A second isoform of the protein called Long Oskar is present in some Dipteran insects, and contains a 5′ domain as well as the three domains of Short Oskar. Below the schematic representation is a rendering of the previously reported solved structures for the LOTUS (PDBID: 5NT7) and OSK (PDBID: 5A4A) domains (Jeske et al. 2015; Yang et al. 2015) with a speculative rendering of the unfolded interdomain region shown with a dashed line.

NCBI insect transcriptomes and genomes that were publicly available at the time of analysis (supplementary table S1, Supplementary Material online; fig. 2; see Genome and Transcriptome Preprocessing in Materials and Methods for NCBI accession numbers and additional information). First, we used the HMMER tool suite to build HMM models for each of the LOTUS and OSK domains, using our previously generated multiple sequence alignments (MSA) (Blondel et al. 2020). We subjected genomes to in silico gene model inference using Augustus (Stanke et al. 2006). We translated the resulting predicted transcripts, as well as the predicted transcripts from RNA-seq data sets, in all six frames. We then scanned the resulting protein sequences for the presence of LOTUS and OSK domains using the aforementioned HMM models. Sequences were designated as oskar homologs based on the same criteria as in our previous study (Blondel et al. 2020), namely, sequences containing both a LOTUS and an OSK domain (Jeske et al. 2015), separated by a variable interdomain region. We then aligned all sequences using hmmalign and the HMM derived from our previously published full length Oskar alignment (Blondel et al. 2020). The first iteration of the alignment was manually curated as previously described (Blondel et al. 2020), and sequence duplicates and sequences that did not align correctly were removed. All subsequent iterations were automatically curated following the process described in Materials and Methods: Identification of Oskar Homologs.
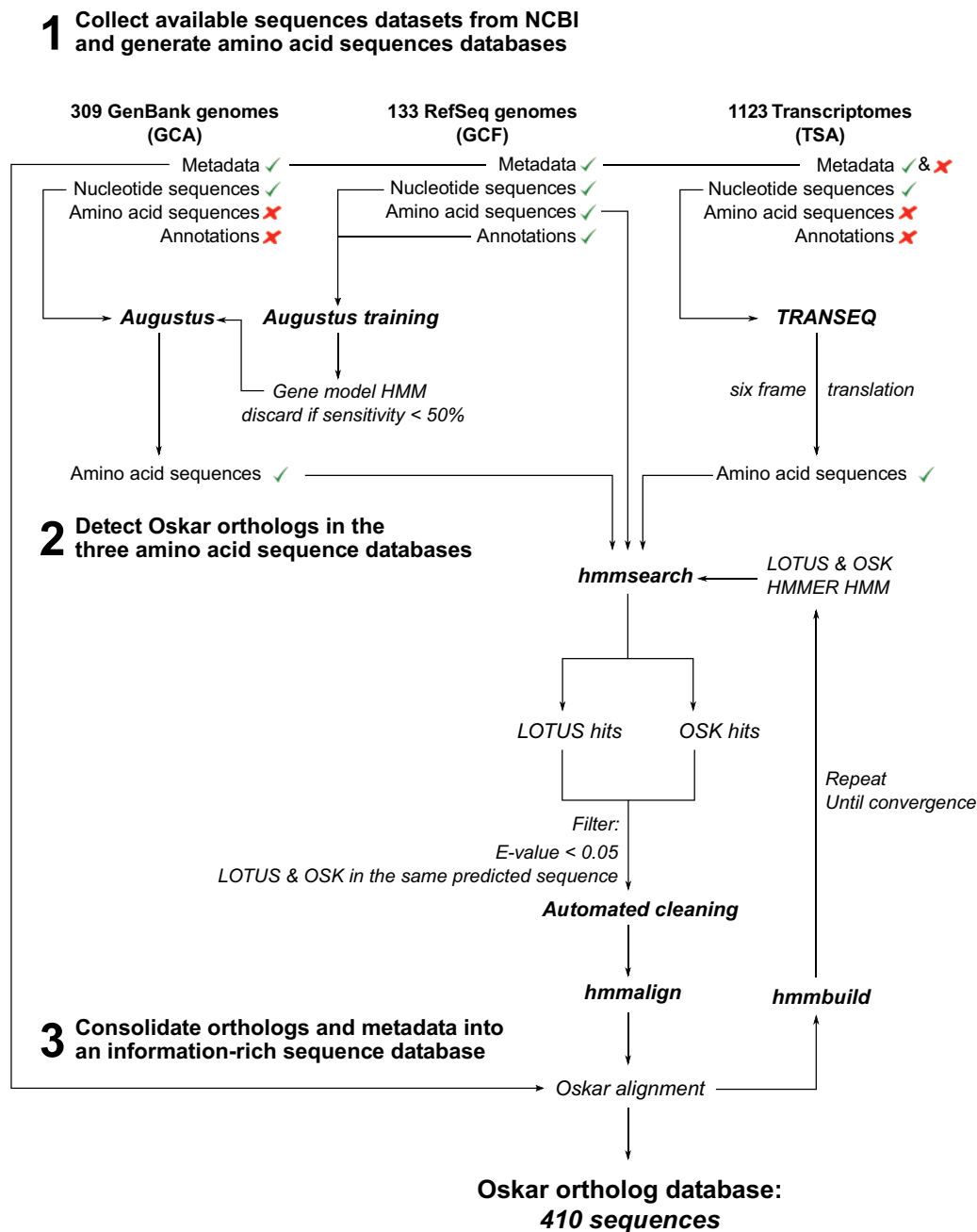
With these methods, we recovered a total of 379 unique oskar sequences from 350 unique species. To our knowledge, this comprises the largest collection of oskar homologs described to date. To determine if oskar homologs might predate Insecta, we applied the discovery pipeline to all 31 genomes and 266 transcriptomes of noninsect pancrustaceans available at the time of analysis (see Genomes and Transcriptomes Preprocessing in Materials and Methods for complete list). However, we did not recover any noninsect sequences meeting our criteria for oskar homologs (fig. 3),

strongly suggesting that oskar is restricted to the insect lineage (Lynch et al. 2011; Ahuja and Extavour 2014).

We found that 58.65% of RefSeq genomes (78/133), 30.42% of GenBank genomes (94/309), and 21.19% of transcriptomes (238/1,123) analyzed contained predicted oskar homologs (supplementary table S1 and fig. S1a, Supplementary Material online). Given that detection of putative homologs is highly dependent on the quality of the genome assembly and annotation, we asked whether there were differences in the assembly statistics of genomes with and without predicted oskar homologs. We observed a significant difference in N50, L50, number of contigs, and number of scaffolds between genomes lacking oskar hits and those where oskar was identified (Mann−Whitney U test P value < 0.05). Genomes where we did not find oskar showed a significantly higher mean/median contig and scaffold count, smaller contig and scaffold N50 length, larger contig and scaffold L50, and more contigs or scaffolds per genome length, than genomes where we detected an oskar homolog (Mann−Whitney U test $P < 0.05$; supplementary fig. S2 and table S2, Supplementary Material online). We interpret this to mean that oskar may appear to be absent from these data sets due to potentially incomplete sequencing, suggesting that deeper sequencing in these lineages could possibly reveal additional new oskar homologs in future studies. However, we note that we believe that our analysis provides strong evidence for true oskar loss in at least some lineages, given their very deeply sequenced and well-annotated genomes (e.g., A. mellifera, T. castaneum).

oskar Predates the Divergence of Ametabola and Other Insects

We identified oskar homologs in 15 of the 29 generally recognized insect orders (Misof et al. 2014), including eight holometabolous orders, six hemimetabolous orders, and one ametabolous order (fig. 3). This result is consistent with our previous proposals that oskar predates the origins of the Holometabola (Ewen-Campen et al. 2012; Blondel et al.

1 Collect available sequences datasets from NCBI and generate amino acid sequences databases

| 309 GenBank genomes (GCA) | 133 RefSeq genomes (GCF) | 1123 Transcriptomes (TSA) |

Metadata ✓ — Metadata ✓ — Metadata ✓ & ✗

Nucleotide sequences ✓ Nucleotide sequences ✓ Nucleotide sequences ✓
Amino acid sequences ✗ Amino acid sequences ✓ Amino acid sequences ✗
Annotations ✗ Annotations ✓ Annotations ✗

Augustus ← *Augustus training* *TRANSEQ*

Gene model HMM discard if sensitivity < 50%

six frame | *translation*

Amino acid sequences ✓ Amino acid sequences ✓

2 Detect Oskar orthologs in the three amino acid sequence databases

hmmsearch ← *LOTUS & OSK HMMER HMM*

LOTUS hits   OSK hits

Repeat Until convergence

Filter:
E-value < 0.05
LOTUS & OSK in the same predicted sequence

Automated cleaning

hmmalign   *hmmbuild*

3 Consolidate orthologs and metadata into an information-rich sequence database

Oskar alignment

Oskar ortholog database: 410 sequences

FIG. 2. Schematic presentation of the *oskar* homolog detection pipeline. Sequences were collected automatically from the three NCBI databases, GenBank (GCA), RefSeq (GCF), and TSA database. RefSeq genomes were used to generate Augustus gene model HMMs, which were used to annotate and predict proteins in the nonannotated genomes obtained from GenBank. Transcripts from the TSA database were six-frame translated using TRANSEQ. Amino acid sequences were consolidated into three protein databases. *hmmsearch* from the HMMER tool suite was used to search for LOTUS and OSK hits in those sequences. Sequences with hits for both the LOTUS and OSK domains with an E-value < 0.05 were annotated as *oskar* sequences. Sequences were then cleaned to remove duplicates (sequences with <80% sequence similarity coming from the same organism). The resulting sequences were aligned using *hmmalign*, and the process was repeated until no new sequences were identified. Finally, the sequences were consolidated with the data set metadata into the *oskar* homolog database that was used for all subsequent analyses.

2020). The novel finding of an *oskar* homolog from the silverfish *Atelura formicaria* (Zygentoma) allows us to date back the origin of *oskar* further than previous analyses, to at least 420 Ma (Misof et al. 2014), before the divergence of Ametabola from the remaining insect lineages.

We then explored the distribution of *oskar* sequences across insect phylogeny. Interestingly, we identified multiple lineages where *oskar* appeared to have been lost

independently, including confirming the previously reported (Lynch et al. 2011) losses from the genomes of the red flour beetle *T. castaneum*, the honeybee *A. mellifera*, and the silk moth *B. mori* (fig. 3). Notably, within Lepidoptera we identified *oskar* homologs in only 3 species, despite the fact that we searched 232 available lepidopteran sequence data sets, including 17 well-annotated RefSeq genomes and 135 transcriptomes (fig. 3; supplementary fig. S3, Supplementary Material
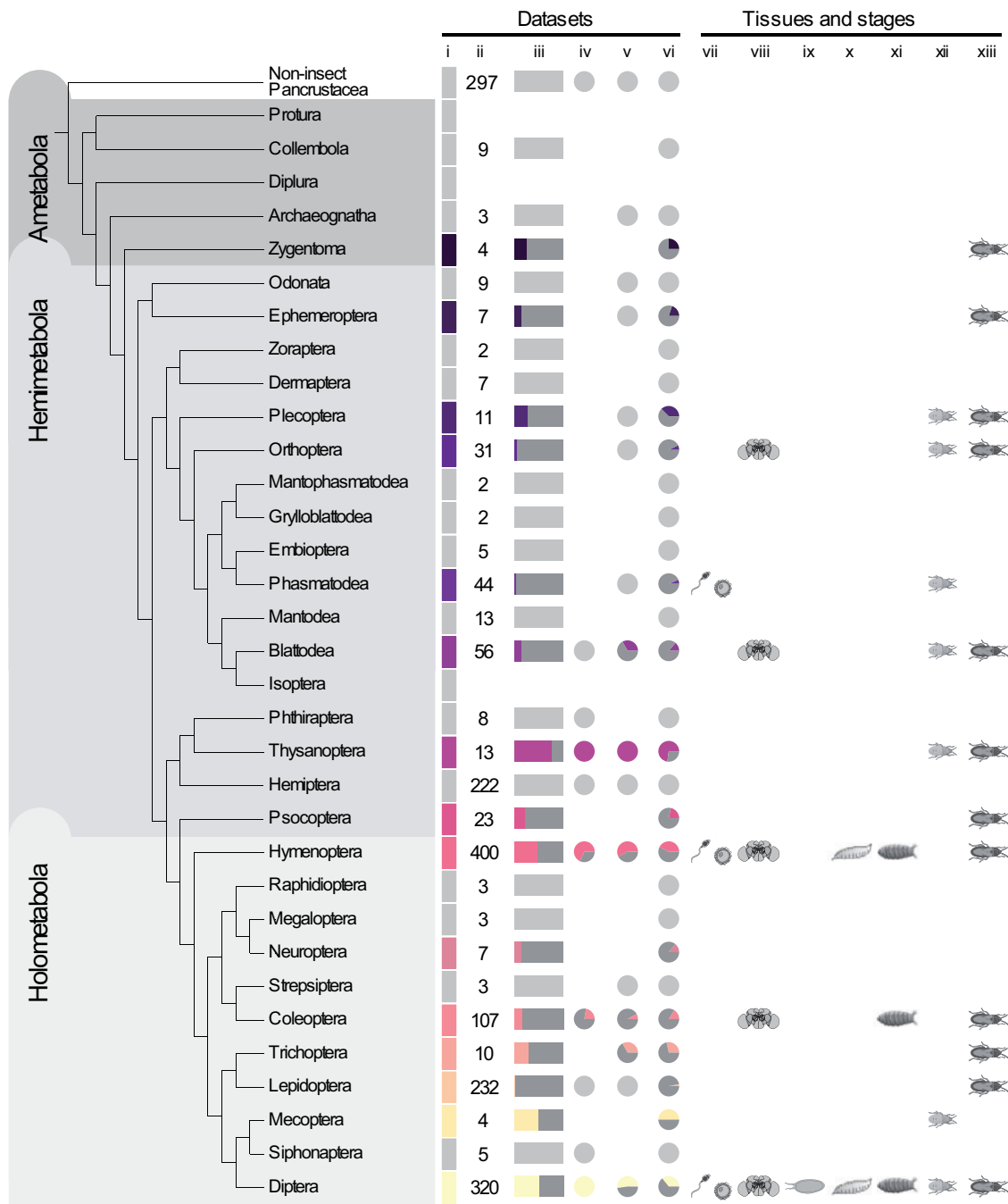
Fig. 3. Summary of *oskar* distribution and expression in insects. Phylogeny from Misof et al. (2014). Symbols in order from left to right: (i) vertical rectangles: gray: no *oskar* homolog was identified in this order. Color (unique for each order): at least one *oskar* homolog was identified in this order. (ii) Number of data sets searched. (iii) Horizontal rectangles: proportion of searched data sets in which an *oskar* homolog was identified. (iv) Pie chart: proportion of *oskar* sequences identified in RefSeq (GCF) data sets. (v) Pie chart: proportion of *oskar* sequences identified in GenBank (GCA) data sets. (vi) Pie chart: proportion of *oskar* sequences identified in TSA database data sets; (vii) *oskar* sequences identified in tissue related to germ line (transcriptomes derived from reproductive organs, eggs, or embryos); (viii) *oskar* sequences identified in tissue related to the brain (transcriptomes derived from brain or head); (ix) *oskar* sequences identified in an egg stage transcriptome; (x) *oskar* sequences identified in a larval stage transcriptome; (xi) *oskar* sequences identified in a pupal stage transcriptome; (xii) *oskar* sequences identified in a nymphal or juvenile stage transcriptome; (xiii) *oskar* sequences identified in an adult transcriptome. All numbers represented graphically here are in supplementary table S1, Supplementary Material online. No data sets were available for Protura, Diplura, or Isoptera at the time of analysis.

online). In principle, this apparent widespread absence of *oskar* in Lepidoptera could be due to unusually rapid evolution of the *oskar* sequence in this lineage, which might render lepidopteran *oskar* homologs undetectable by our methods.

However, we note that the only four lepidopteran homologs we detected all belonged to species of the basally branching *Adelidae* and *Palaephatidae* families. We therefore favor the interpretation that *oskar* was lost from a last common

ancestor of *Meessiidae* and *Palaphaetidae*, ~180 Ma, with the consequence that the majority of extant lepidopteran lineages lack an *oskar* homolog (supplementary fig. S3, Supplementary Material online) (Mitter et al. 2017; Kawahara et al. 2019).

The Hemiptera also appear to have lost *oskar*, based on our analysis of the 222 data sets available for this clade, including 12 RefSeq genomes and 192 transcriptomes. However, we did identify an *oskar* homolog in the Thysanoptera, which is a hemipteran sister group (Misof et al. 2014). Finally, we identified *oskar* homologs in only four of the 11 orders of the Polyneoptera for which data were available. With the exception of Mantodea (13 transcriptomes), the four orders with detectable *oskar* sequences all had more than ten available sequence data sets (Plecoptera: three genomes and eight transcriptomes; Orthoptera: three genomes and 28 transcriptomes; Phasmatodea: 13 genomes and 31 transcriptomes; Blattodea: five genomes and 51 transcriptomes). The remaining orders had fewer than eight data sets each available for analysis (fig. 3; supplementary table S1, Supplementary Material online), which could account for the apparent paucity of *oskar* genes in this group. However, we cannot rule out the possibility that *oskar* in the Polyneoptera may have diverged beyond our ability to detect it, or that it may have been lost multiple times, as observed for multiple holometabolous orders.

As well as multiple convergent losses of *oskar*, we also uncovered evidence for independent instances of duplication of the *oskar* locus. We defined a putative duplication instance as two or more *oskar* sequences (possessing both a LOTUS and OSK domain as per our definition) in the same species that shared <80% sequence similarity. All of these events were detected within the Hymenoptera. We therefore performed a phylogenetic analysis of the hymenopteran sequences to test the hypothesis that these were the result of duplication events (fig. 4; supplementary fig. S4, Supplementary Material online). Our analysis of hymenopteran *oskar* sequences recovered previously published hymenopteran phylogenetic relationships (Peters et al. 2017). We found that *oskar* was duplicated in the four Figitidae species studied, a family of parasitoid wasps. Moreover, one out of ten examined Cynipidae species, as well as the only Ceraphronidae species examined, also harbored a duplicated *oskar* sequence. Multiple *oskar* duplications were also identified in the Chalcidoid wasps, notably in the Mymaridae (all three species studied), the Eupelmidae (two out of three species), the Aphelinidae (both species), and the Pteromalidae (one out of 17 species). Finally, we identified two additional apparently independent duplication events in the Aculeata, one in the wasp *Polistes fuscatus* [of 29 Vespidae, including three additional *Polistes* species, two with RefSeq genomes (*P. canadensis* and *P. dominula*) in which *oskar* was identified in single copy], and one in the red imported fire ant *Solenopsis invicta* (of 41 Formicidae species, including the congeneric *S. fugax*, with a GenBank genome in which *oskar* was identified in single copy).

Evidence for *oskar* Expression in Multiple Somatic Tissues

In studied insects to date, *oskar* is expressed and required in one or both of the germ line (Juhn and James 2006; Juhn et al. 2008; Lynch et al. 2011; Lehmann 2016) or the nervous system (Ewen-Campen et al. 2012; Xu et al. 2013). We asked whether these expression patterns could be detected in the insects studied here. To this end, we downloaded all available metadata for the transcriptomes analyzed here, to obtain information on the source tissues and developmental stages. We obtained these data for 371 out of the 1,123 transcriptomes in our analysis, including both holometabolous and hemimetabolous orders (see TSA Metadata Parsing and Curation in Materials and Methods). To first explore the distribution of *oskar* expression in the brain and the germ line, we binned the different tissues reported in the metadata into two categories, brain or germ line. This was done independently of the developmental stage (if that information was included in the metadata) by creating a mapping table and checking the extracted tissues against this table (supplementary table S3 at GitHub repository TableS3_germline_brain_table.csv, Supplementary Material online). We then cross referenced our homology detection with these metadata. We found evidence for *oskar* expression in the germ line of four orders (Phasmatodea, Hymenoptera, Coleoptera, and Diptera), and in the brain of five orders (Orthoptera, Blattodea, Hymenoptera, Coleoptera, Diptera) (see TSA Metadata Parsing and Curation in Materials and Methods for details on keyword extractions). For the vast majority of the data sets examined, transcriptomes were not generated with comparable methods for different organ systems from the same species, such that we cannot make strong statements about the relative expression levels of *oskar* in the reproductive and nervous systems. However, we did perform a limited assessment of this question using previously published transcriptomes from the mosquito *Aedes aegypti* (Diptera) (Matthews et al. 2016) (supplementary fig. S5, Supplementary Material online) and the cricket *Gryllus bimaculatus* (Orthoptera) (Whittle, Kulkarni, Chung, et al. 2021; Whittle, Kulkarni, and Extavour 2021) (supplementary fig. S6, Supplementary Material online), and RT-PCR on isolated gonads and heads from *D. melanogaster* (Diptera), the weevil *Callosobruchus maculatus* (Coleoptera), and the stick insect *Aretaon asperrimus* (Phasmatodea) (supplementary materials, Supplementary Material online). For *D. melanogaster*, significant expression was detected only in female gonads (supplementary fig. S7, Supplementary Material online). For the remaining four species, whereas *oskar* transcripts were detected in both gonads and heads, levels appeared higher in gonads than in heads (supplementary figs. S5—S7, Supplementary Material online).

In addition, we found evidence of *oskar* expression in several somatic tissues not previously implicated in studies of *oskar* expression and function. These tissues included the midgut (*P. fuscatus*, *Sitophilus oryzae*), fat body (*P. fuscatus*, *Arachnocampa luminosa*), salivary gland (*Culex tarsalis*, *Anopheles aquasalis*, *Leptinotarsa decemlineata*), venom gland (*Culicoides sonorensis*, *Fopius arisanus*), and silk gland (*Bactrocera cucurbitae*)
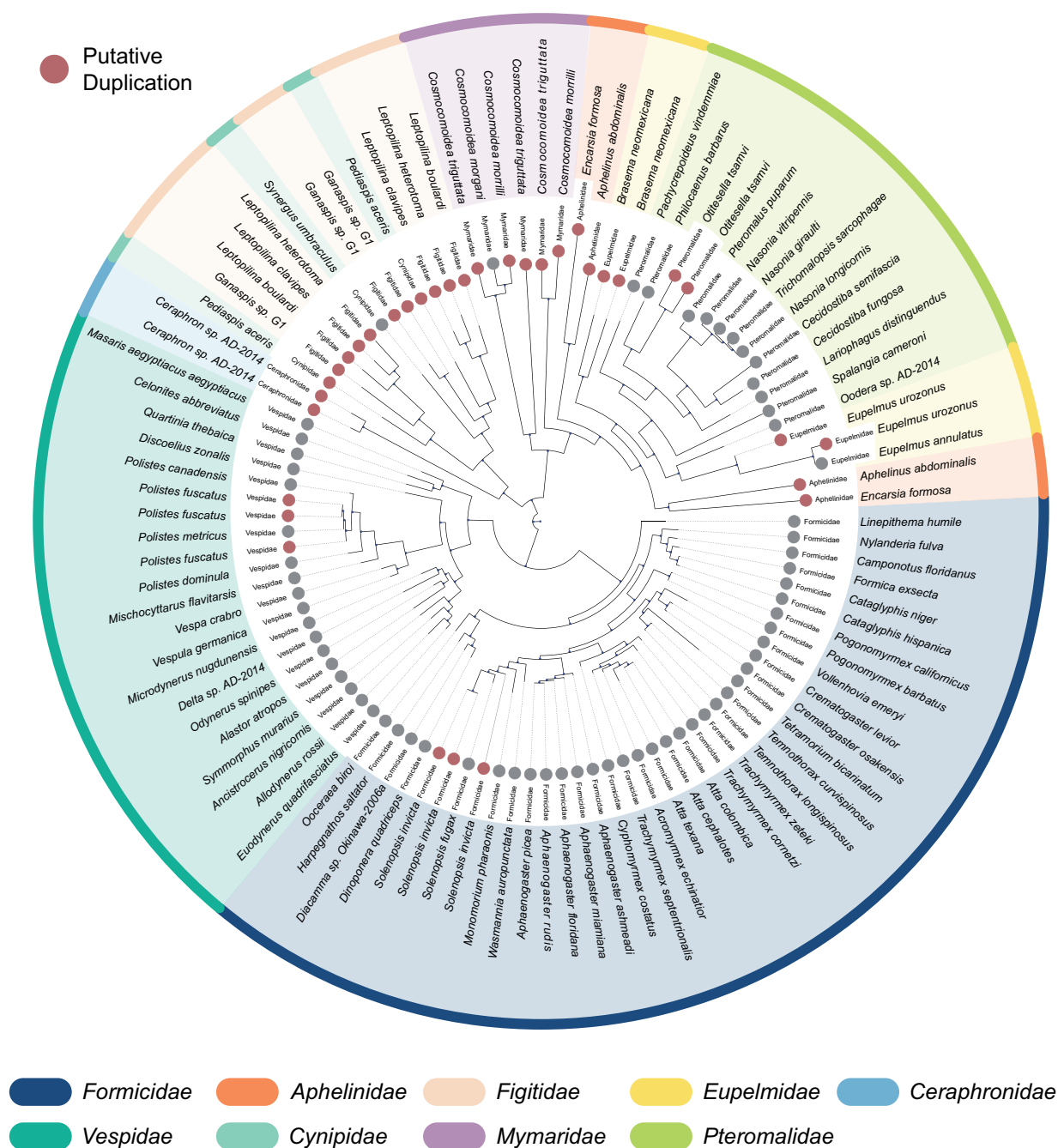
FIG. 4. Phylogenetic reconstruction of hymenopteran Oskar sequences. Phylogenetic tree inferred using RaxML with 100 bootstraps. Each leaf represents an Oskar homolog. Gray circles: Only one Oskar sequence was identified. Red circles: putatively duplicated Oskar sequences identified (sequence similarity <80%). Only families which contained a putative duplication are shown here; see supplementary figure S4, Supplementary Material online, for the results of our *oskar* search in the context of a more complete hymenopteran phylogeny.

(supplementary fig. S8, Supplementary Material online). In terms of developmental stage, we detected expression of *oskar* during embryonic, larval, or nymphal stages only in holometabolous insects, and for most hemimetabolous insects, *oskar* was detected in transcriptomes derived from adults (fig. 3). However, it is important to note that for most species, transcriptomes were available only from adult tissues, rather than from a full range of developmental stages (supplementary fig. S5, Supplementary Material online). We therefore cannot rule out the possibility that *oskar* expression at preadult stages is

also a feature of multiple Hemimetabola. Indeed, we previously reported that *oskar* is expressed and required in the embryonic nervous system of a cricket, a hemimetabolous insect (Ewen-Campen et al. 2012).

The Long *oskar* Domain is an Evolutionary Novelty Specific to a Subset of Diptera

Drosophila melanogaster has two isoforms of Oskar (Markussen et al. 1995): Short Oskar, containing the LOTUS, OSK and interdomain regions, and Long Oskar,

containing all three domains of Short Oskar as well as an additional 5' domain (supplementary fig. S9, Supplementary Material online). It was previously reported that Long Oskar was absent from *N. vitripennis*, *C. pipiens*, and *G. bimaculatus* (Lynch et al. 2011; Ewen-Campen et al. 2012), and within our alignment of Oskar sequences we could only detect the Long Oskar isoform within Diptera. Therefore, using our data set, we asked when these two isoforms had evolved. We selected the dipteran sequences from our Oskar alignment and then grouped the sequences by family. We plotted the amino acid occupancy at each alignment position (supplementary fig. S9, Supplementary Material online), and found that Long Oskar predates the Drosophilids, being identified as early as the *Pinpunculidae* (supplementary fig. S9, Supplementary Material online). Moreover, following the evolution of the Long Oskar isoform, the Long Oskar domain was retained in all families except for the *Glossinidae* and *Scathophagidae*. However, given that we identified only eight and two Oskar sequences for these families respectively, we cannot eliminate the possibility that apparent absence of the Long Oskar domain in these groups reflects our small sample size, rather than true evolutionary loss.

The LOTUS and OSK Domains Evolved Differently between Hemimetabolous and Holometabolous Insects

The fact that an *oskar*-dependent germ plasm mode of germ line specification mechanism has been identified only in holometabolous insects suggests that *oskar* may have been co-opted in this clade for this function (Ewen-Campen et al. 2012). Under this hypothesis, evolution of the *oskar* sequence in the lineage leading to the Holometabola may have changed the physico-chemical properties of Oskar protein, such that it acquired germ plasm nucleation abilities in these insects. To test this hypothesis, we asked whether there were particular sequence features associated with Oskar proteins from holometabolous insects, in which Oskar can assemble germ plasm, and hemimetabolous insects, which appear to lack *oskar*-dependent germ plasm. In particular, we assessed the differential conservation of amino acids at particular positions across Oskar and asked if these might be predicted to change the physico-chemical properties of Oskar in specific ways that could potentially be relevant to germ plasm nucleation. We used the Valdar score (Valdar 2002) as the main conservation indicator for this study (see GitHub file scores.csv), as this metric accounts not only for transition probabilities, stereochemical properties and amino acid frequency gaps, but also for the availability of sequence diversity in the data set. It computes a weighted score, where sequences from less well-represented clades contribute proportionally more to the score than sequences from over-represented clades. Due to the highly unbalanced availability of genomic and transcriptomic data between hemimetabolous and holometabolous sequences (supplementary table S1, Supplementary Material online; fig. 3) the choice of a weighted score was necessary to avoid biasing the results toward insect orders such as Diptera or Hymenoptera. To study the difference

between hemimetabolous and holometabolous sequences, we did not use the Valdar score directly, but instead computed the conservation ratio between both groups for each position, which we call the conservation bias (see Computation of the Conservation Bias in Materials and Methods). We plotted the conservation bias on the solved 3D crystal structure of the *D. melanogaster* LOTUS and OSK domains (Jeske et al. 2015; Yang et al. 2015) to ask whether specific functionally relevant structures showed phylogenetic or other patterns of residue conservation (fig. 5).

First, we asked if the conservation score at the scale of domains was different between holometabolous and hemimetabolous sequences. We observed that the conservation bias for the LOTUS domain was centered around a mean of 1.00, indicating that both Holometabola and Hemimetabola displayed a similar conservation of the LOTUS domain (fig. 5a). For the OSK domain, however, the conservation bias was centered around 0.84, indicating that the hemimetabolous sequences displayed a higher level of conservation compared with holometabolous sequences (fig. 5a). To interrogate specific biochemical hypotheses, we then examined the degree of conservation bias in different regions of the protein structure. We asked if the amino acids of the β sheets of the LOTUS domain thought to be involved in dimerization of the protein (Jeske et al. 2015; Yang et al. 2015) displayed conservation bias. Both β sheets had an overall even bias (mean: 1.03 and 1.05 for β1 and β2, respectively) between both groups (fig. 5b). Second, as we had observed that hemimetabolous OSK was more conserved overall than holometabolous OSK, we asked if there were any clear patterns of conservation bias in specific regions of the OSK domain (fig. 5a and b). We found that some of the secondary structures within the OSK domain showed a differential conservation (α2: 0.54, α6: 0.42, β2: 0.52), whereas other structures were within <0.1 of the median value for OSK. Moreover, we observed a large pocket of amino acids showing a conservation bias toward hemimetabolous sequences located on the surface of OSK (fig. 5c). This particular area contains the previously reported important amino acids for the RNA binding function of OSK (Jeske et al. 2015; Yang et al. 2015) namely, R442, R436, and R576. The electrostatic properties at those positions were conserved in the holometabolous sequences R436: 0.36, R442: 0.29 and R576: 0.81 (fig. 5d), but not in hemimetabolous sequences. In other words, these specific amino acid residues are outliers in that they are more specifically conserved in holometabolous OSK sequences, but are located within a domain that overall is more conserved in Hemimetabola.

To gain further insight into the differences in conservation across insects, we reduced the MSA dimensionality using a multiple correspondence analysis (MCA), an equivalent of PCA for categorical variables (Lebart et al. 1984). We performed the dimensionality reduction for the full-length Oskar sequence alignment as well as for the LOTUS and OSK alignments (supplementary fig. S10, Supplementary Material online). Interestingly, we found that most of the variance in sequence space was due to dipterans and hymenopterans (supplementary fig. S10, Supplementary Material online). When we considered the OSK domain only, we
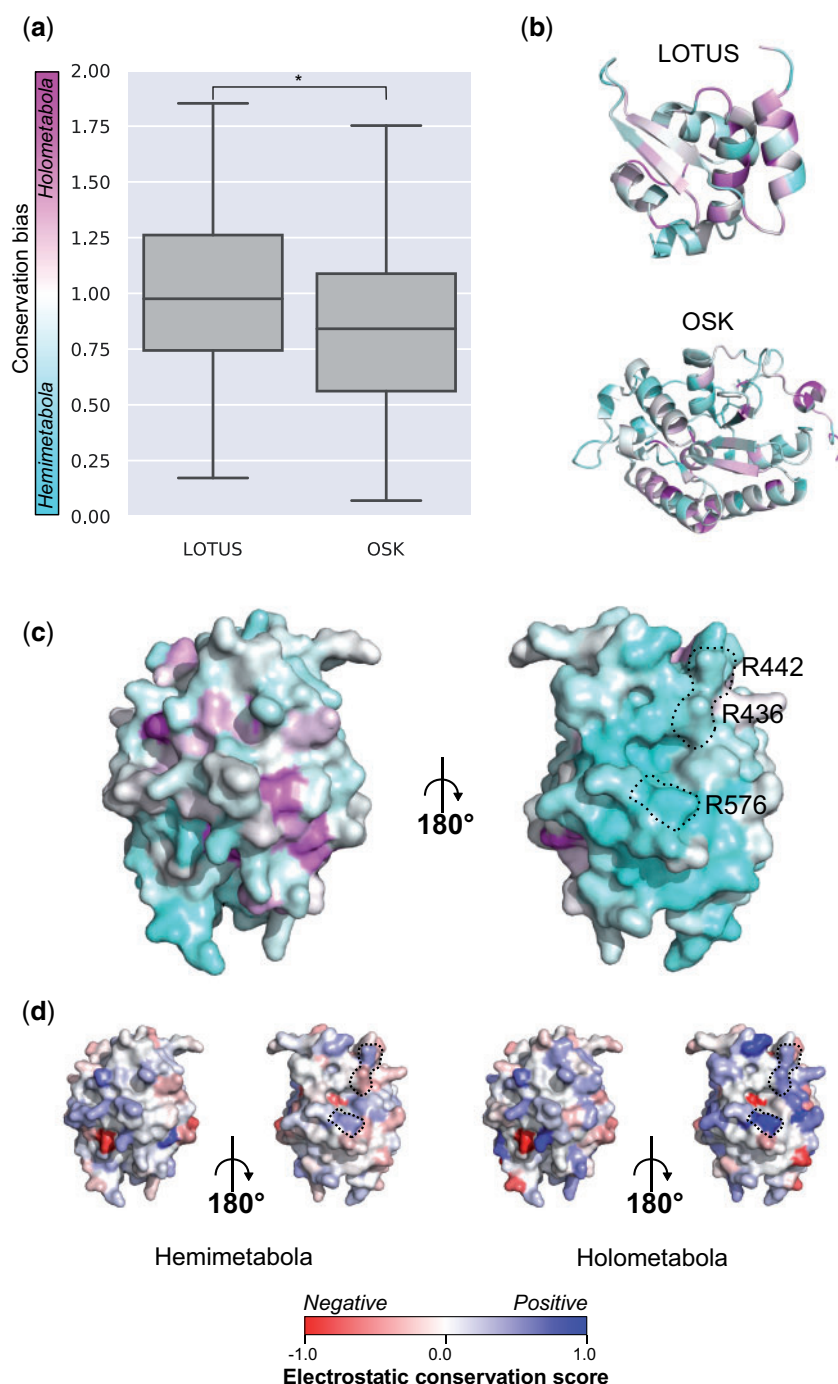
FIG. 5. Differential conservation of amino acids between hemimetabolous and holometabolous Oskar sequences. (*a*) Box plot showing the conservation bias for each of the LOTUS and OSK domains between hemimetabolous and holometabolous Oskar sequences. Statistical difference was tested using a Mann−Whitney U test ($P < 0.05$). (*b*) Ribbon diagram of LOTUS (PDBID: 5NT7) and OSK (PDBID: 5A4A) domain structures, where each amino acid is colored by conservation bias on the color scale shown in (*a*). (*c*, *d*) Protein surface representation of the OSK domain (PDBID: 5A4A) from two different angles. Black dashed lines indicate the three amino acids reported previously to be necessary for OSK binding to RNA in *D. melanogaster* (Jeske et al. 2015; Yang et al. 2015). (*c*) Amino acids colored by conservation bias on the color scale shown in (*a*). Cyan: amino acids more highly conserved in hemimetabolous sequences; magenta: amino acids more highly conserved in holometabolous sequences. (*d*) Amino acids colored by electrostatic conservation score. Left: hemimetabolous sequences; right: holometabolous sequences.

identified clusters of *Drosophilidae*, *Culicidae*, and *Formicidae* sequences (supplementary fig. S10, Supplementary Material online). This clustering is also reflected for the LOTUS

domain, where the *Drosophilidae* and *Culicidae* contribute to a high amount of variance in the first MCA dimension. However, for the LOTUS domain, the *Formicidae* sequences

do not cluster away from other Oskar sequences (supplementary fig. S10, Supplementary Material online). This suggests that the LOTUS domain of Diptera diverged in sequence between *Drosophilidae* and *Culicidae*.

Evidence for Evolution of Stronger Dimerization Potential of the Oskar LOTUS Domain in Holometabola

The LOTUS domain dimerizes in vitro through electrostatic and hydrophobic contacts of Arg215 of the β2 sheet and Thr195, Asp197, and Leu200 of the α2 helix (Jeske et al. 2015; Yang et al. 2015). To date, however, the biological significance of Oskar dimerization remains unknown. Moreover, the dimerization of the LOTUS domain does not appear to be conserved across all Oskar sequences (Jeske et al. 2015). Specifically, ten LOTUS domains from nondrosophilid species were tested for dimerization, and only LOTUS domains from *Drosophilidae*, *Tephritidae*, and *Pteromalidae* formed homodimers (Jeske et al. 2015). The other sequences tested, from *Culicidae*, *Formicidae*, and *Gryllidae*, remained monomeric under the tested conditions (Jeske et al. 2015). We selected the LOTUS sequences in our alignment from those six families and placed them into one of two groups, dimeric and monomeric LOTUS, under the assumption that any sequence from that family would conserve the dimerization (or absence thereof) properties previously reported (Jeske et al. 2015). We asked whether we could detect any evolutionary changes between the two groups in properties of known important dimerization interfaces and residues in our sequence alignment (Jeske et al. 2015).

In the *D. melanogaster* structure, two key amino acids, D197 and R215, are predicted to form hydrogen bonds that stabilize the dimer (Jeske et al. 2015). We found that in the dimer group, the electrostatic properties of these two amino acids are highly conserved (−0.75 for D197 and 0.81 for R215), whereas in the monomer group the electrostatic interaction is not conserved (0.03 for D197 and −0.11 for R215) (fig. 6e). Given the differential conservation between the two groups, our results support the previous finding that disrupting this interaction prevents dimerization (Jeske et al. 2015). L200 was previously hypothesized to stabilize the interface via hydrophobic forces (Jeske et al. 2015). We observed that the hydrophobicity of this residue is highly conserved in the dimer group (L200: 0.89), but that in the monomer group this residue is hydrophilic (L200: 2.33) (fig. 6f). In sum, our analyses show that key amino acids in the LOTUS domain evolved differently in distinct insect lineages, in a way that may explain why some insect LOTUS domains dimerize and some do not.

Conservation of the Oskar−Vasa Interaction Interface

Next, we asked whether we could detect differential conservation of regions or residues within the LOTUS−Vasa interface. It was previously reported that the LOTUS domain of Oskar acts as an interaction domain with Vasa (Jeske et al. 2017), a key protein with a conserved role in the establishment of the animal germ line (Hay et al. 1990; Lasko 2013). The interaction between *D. melanogaster* Oskar's LOTUS

domain and Vasa is through an interaction surface situated in the pocket formed by the helices α2 and α5 of the LOTUS domain (fig. 6a−c). Due to the essential role that *vasa* plays in germ line determination (reviewed in Raz [2000]; Noce et al. [2001]; Extavour and Akam [2003]; Ewen-Campen et al. [2010]; Lasko [2013]), and the potential co-option of *oskar* to the germ line determination mechanism in Holometabola (Ewen-Campen et al. 2012), we hypothesized that evolutionary conservation of the residues of this interface might be detectable. First, we observed that the residues of the LOTUS domain α2 and α5 helices, which directly contact Vasa (Jeske et al. 2017) were highly conserved overall (α2 average Valdar score 0.49; α5 Valdar score 0.56) (fig. 6b). Specifically, we observed that the previously in vitro-confirmed Vasa interacting amino acids A162 and L228 of the LOTUS domain were highly conserved (Valdar score: 0.64 for both residues) (Jeske et al. 2017). We also noted that Q235 and H227 of the LOTUS domain α5 helix are also highly conserved, suggesting them as putative novel important interaction partners (Valdar score: 0.90 and 0.90 for both residues) (fig. 6b). Moreover, facing the LOTUS domain H227 is Vasa M540, which may act as a proton donor to form a hydrogen bond between the histidine ring and the sulfur atom of the methionine (Pal and Chakrabarti 2001) (fig. 6b and b'). The LOTUS domain α2 helix is overall slightly less conserved than the LOTUS domain α5 helix (Valdar score: 0.49 vs. 0.56) (fig. 6a, b'', and c''), but hydrophobic properties are conserved on one side of the α2 helix (fig. 6c and c') forming a motif of conserved amino acid properties (fig. 6c'').

Previous reports have hypothesized that the *D. melanogaster* LOTUS domain could act as a dsRNA binding domain (Anantharaman et al. 2010; Callebaut and Mornon 2010). However, in *D. melanogaster*, it was later reported that the LOTUS domain did not bind to nucleotides (Jeske et al. 2015). Therefore, using our data set we assessed the potential RNA binding properties of LOTUS domains to test the conservation of this prediction. We used the RNABindR algorithm (Terribilini et al. 2007) to predict potential RNA binding sites of the LOTUS domain, and computed a conservation score for each position (Terribilini et al. 2007). We found that the α5 helix is the location in the LOTUS domain that has the most conserved prediction for RNA binding (fig. 6d). We therefore suggest that the possibility that LOTUS binds RNA directly warrants further experimental examination.

Finally, we asked whether the secondary structure of the LOTUS domain might be conserved. Secondary structures are often indicative of the tertiary structure of a domain. Therefore, we reasoned that the secondary structure might be conserved even if the sequence varies. We submitted the LOTUS sequences from all identified Oskar homologs to the Jpred4 servers (Drozdetskiy et al. 2015) for secondary structure prediction and mapped the results onto the Oskar alignment we obtained. We found that the secondary structure of LOTUS is highly conserved throughout Oskar homologs, with the exception of the α1 helix (supplementary fig. S11, Supplementary Material online) which displays a low conservation score of 0.19 (fig. 6a).
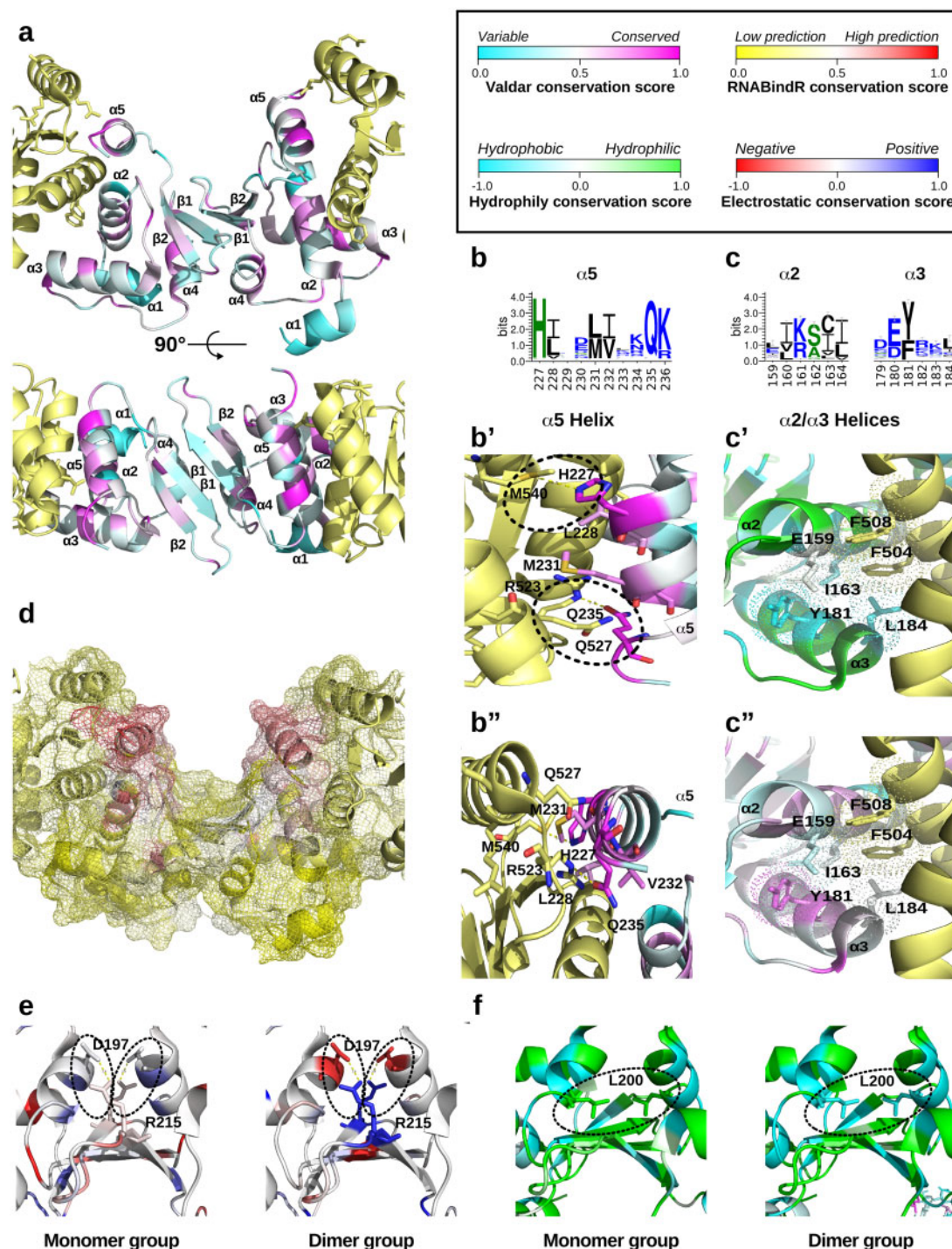
FIG. 6. Conservation analysis of the LOTUS domain. (*a*) Ribbon diagram of a LOTUS domain dimer (cyan/magenta) in complex with two Vasa molecules (yellow) (PDBID: 5NT7) from two different angles. Each LOTUS amino acid is colored based on its Valdar conservation score. (*b, c*) Sequence Logo of the α5 and α2/α3 helices, respectively, generated with WebLogo (Crooks et al. 2004). Black: hydrophobic residues; blue: charged residues; green: polar residues. (*b', b"*) Ribbon diagram of the conserved α5 helix, with key amino acids displayed as sticks and colored by Valdar conservation score. Two potential novel Vasa-LOTUS contacts (H227 and Q235) are highlighted with dashed lines. (*c'*) Ribbon diagram of the conserved α2 helix, with key amino acids displayed as sticks and colored by hydrophobicity/hydrophily conservation score. (*c"*) Ribbon diagram of the conserved α2 helix, with key amino acids displayed as sticks and colored by Valdar conservation score. (*d*) Surface mesh rendering colored with the RNABindR RNA binding conservation score. (*e, f*) Ribbon diagram of the LOTUS β sheet dimerization interface. Left: conservation of monomeric LOTUS domains; right: dimeric LOTUS domains. (*e*) Amino acids colored by electrostatic conservation score. Dashed lines indicate the key electrostatic interaction thought to stabilize the dimerization. (*f*) Amino acids colored by hydrophobicity/hydrophily conservation score. Dashed lines indicate the key hydrophobic pocket thought to stabilize the dimerization.

The Core of the OSK Domain Is Conserved

We asked whether the OSK domain showed any differential conservation across the different parts of the domain. We found that the OSK domain of Oskar showed an overall conservation across all insects, similar to the LOTUS domain (Valdar score: 0.51) (fig. 7a). However, the conservation pattern is higher in the core amino acids (Valdar score average of core amino acid: 0.54) when compared with the residues at the surface (Valdar score average for surface amino acid: 0.23) (fig. 7a). Despite the overall low conservation of the residues at the surface of the OSK domain, we found that the electrostatic properties are conserved overall (electrostatic conservation score > 0; conserved) in the previously reported putative RNA binding pocket (Yang et al. 2015). However, as previously mentioned, this conservation is stronger in holometabolous sequences (fig. 5d). These results are in accordance with the potential role of OSK as an RNA Binding domain in the context of germ plasm assembly (Jeske et al. 2015; Yang et al. 2015). We also submitted the OSK sequences to the same secondary structure analysis performed on LOTUS. We found that, as for the LOTUS domain, the secondary structure of OSK is highly conserved throughout all insect sequences analyzed (supplementary fig. S11, Supplementary Material online).

We then asked if the conservation patterns observed at the core of OSK were clustered in sequence motifs. When we looked at the location of the highly conserved amino acids, we found that the conservation was driven by four well-defined sequence motifs (fig. 7c, c′, c″, and c‴). Given that oskar plays different roles in Holometabola and Hemimetabola, we asked whether the conserved OSK motifs showed any difference in conservation between these two groups. Of the four highly conserved OSK core motifs (fig. 7c, c′, c″, and c‴), two of them (fig. 7c: Valdar average score: 0.80 and fig. 7c″: Valdar average score: 0.71) were conserved across all insects, but the other two showed differential conservation between the holometabolous and hemimetabolous sequences (fig. 7c′: Valdar score average Holometabola: 0.78, Hemimetabola: 0.58; and fig. 7c″: Valdar score average Holometabola: 0.70, Hemimetabola: 0.55). Finally, we noted that only one of the affected OSK domain residues in known loss of function oskar alleles affecting posterior patterning in D. melanogaster, S457, is conserved across all insects (Valdar score: 0.86). This suggests that the role of the other previously reported important amino acids in the function of D. melanogaster OSK (Yang et al. 2015) might not be conserved in other insects (red positions in fig. 7c, c′, c″, and c‴).

## Discussion

An Expanded Collection of oskar Homologs

oskar provides a powerful case study of functional evolution of a gene with an unusual genesis (Blondel et al. 2020). Here, we gathered the most extensive set of homologous oskar sequences to date. However, most insect genomic and transcriptomic data have been generated from only a few orders, and the vast majority from the Holometabola. Diptera, Lepidoptera, Coleoptera, Hymenoptera, and Hemiptera represent 82% of the data sets available at the time of this analysis. We emphasize that expanded taxon sampling, particularly for the Hemimetabola, will be critical for further studies of the evolution of protein function across insects. Moreover, only a small proportion (27% for tissue type, 26% for organism stage, and 14% for sex) of the TSA data sets contained usable metadata regarding the stage and tissue type sampled. Standardization of the nature and format of transcriptomic metadata would also be a worthwhile endeavor that could increase the efficiency and efficacy of future work.

Convergent Losses and Duplications of oskar in Insect Evolution

A previous report suggested that oskar had been lost from the genome of the silk moth B. mori (Lynch et al. 2011). Our analysis of 232 data sets across 44 of the 126 described lepidopteran families (Kawahara et al. 2019) strongly suggests that the loss of oskar in the Lepidoptera (butterflies and moths) is not unique to the silk moth, but rather occurred early and repeatedly in lepidopteran evolution. The fact that oskar is a component of the oosome at the posterior of the oocyte (the wasp germ plasm analog; Quan et al. 2019) and required for germ cell formation in the wasp Nasonia vitripennis (Lynch et al. 2011) implies that a common ancestor of Holometabola had already established an oskar-dependent inheritance mode of germ line specification. Therefore, the apparent subsequent loss in nearly all Lepidoptera examined of a gene responsible for the establishment of the germ plasm in other Holometabola might seem unexpected. Few studies have directly addressed the molecular mechanisms of germ cell specification in Lepidoptera. In B. mori (Bombicidae), vasa mRNA (Nakao 1999), and protein (Nakao et al. 2006), and the transcripts of one of four nanos homologs (nanos-O) (Nakao et al. 2008), have been detected in a region of ventral cortical cytoplasm in preblastoderm stage embryos. As putative primordial germ cells form in this location at later stages (Miya 1958), some authors have speculated that a germ plasm, located ventrally rather than posteriorly, may specify germ cells in this moth (Toshiki et al. 2000; Nakao et al. 2008). However, recent knockdown experiments showed that maternal nanos-O is dispensable for germ cell formation (Nakao and Takasu 2019), consistent with a zygotic, inductive mechanism. In the butterfly Pararge aegeria (Nymphalidae), no oskar homolog has been identified in the genome (Carter et al. 2013), but the transcripts of one of four identified nanos homologs (nanos-O) have been detected in a small region of ventral cortical ooplasm, again prompting speculation that this lepidopteran may also deploy a germ plasm (Carter et al. 2015). We suggest that if these or other Lepidoptera do indeed rely on germ plasm to specify their germ line, they may do so using a germ plasm nucleator other than Oskar. For most studied Lepidoptera, however, classical embryological studies report the first appearance of primordial germ cells at postblastoderm stages, either from the ventral midline of the cellular blastoderm or early germ band (Woodworth 1889; Tomaya 1902; Sehl 1931; Miya 1953, 1958, 1975; Tanaka 1987), from the celomic sac mesoderm
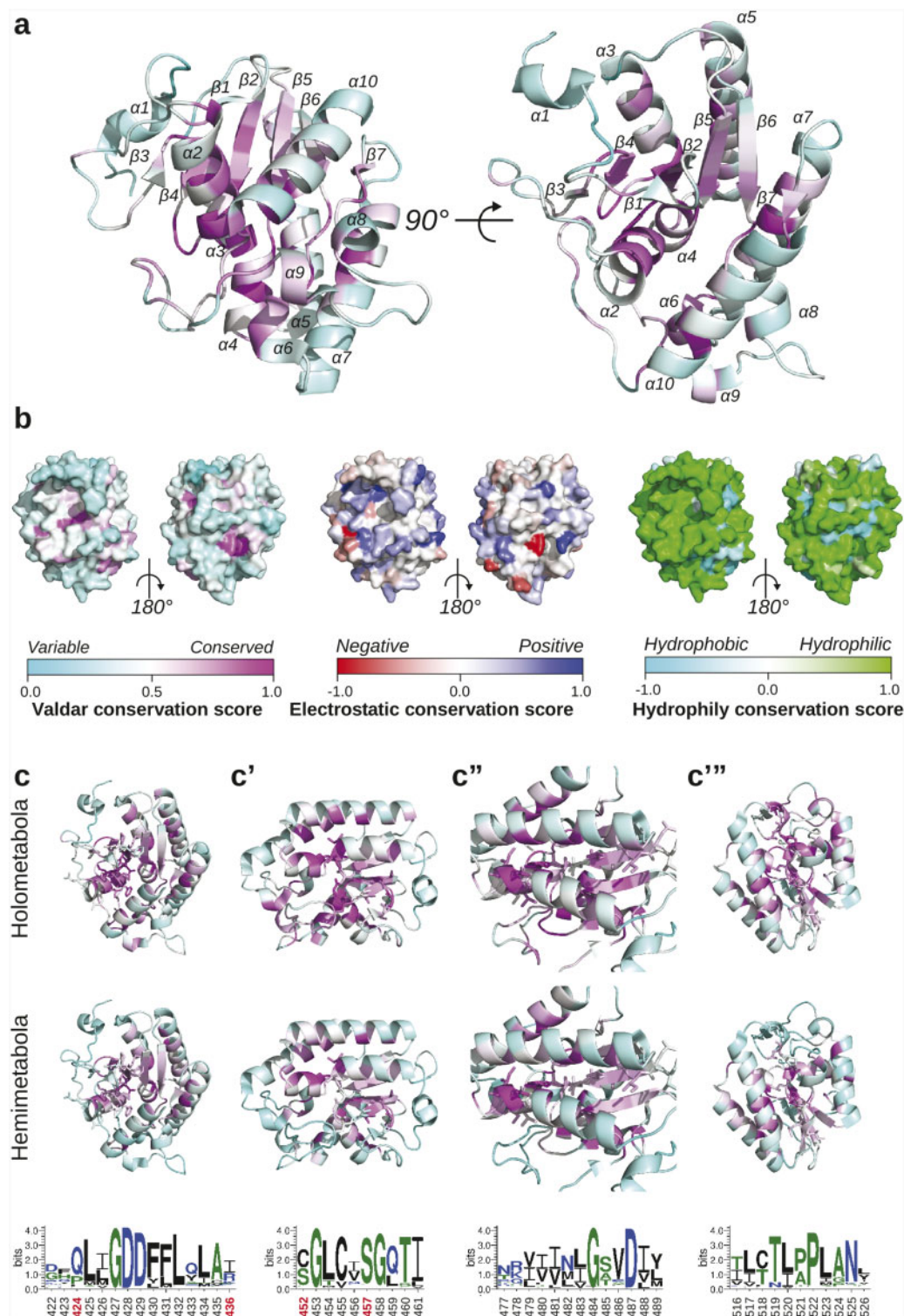
FIG. 7. Conservation analysis of the OSK domain. (*a*) Ribbon diagram of the OSK domain (PDBID: 5A4A) from two different angles. Each amino acid is colored based on its Valdar conservation score. (*b*) Protein surface representation of the OSK domain colored by Valdar conservation, electrostatic conservation and hydrophobicity/hydrophily conservation score. (c, c', c'', c''') Ribbon diagram of newly detected conserved motifs of the OSK domain, showing sequence Logo (bottom row) residues as sticks. Each amino acid is colored with Valdar conservation scores of holometabolous (top row) and hemimetabolous (middle row) OSK sequences. Bottom row: sequence Logos of each conserved motif generated with WebLogo (Crooks et al. 2004). Black: hydrophobic residues; blue: charged residues; green: polar residues. Red numbers: amino acid locations of *D. melanogaster* loss of function *oskar* alleles leading to the loss of *oskar* localization to the posterior pole during embryogenesis (P425S = *osk[8]* (Kim-Ha et al. 1991); S452L = *osk[255]* = *osk[7]* (Lehmann and Nüsslein-Volhard 1986; Kim-Ha et al. 1991); S457F = *osk[6B10]* (Breitwieser et al. 1996)) or to reduced RNA-binding affinity of the OSK domain (R436E; Yang et al. 2015).

of the abdomen (Johannsen 1929; Eastham 1930; Saito 1937; Presser and Rutschky 1957; Kobayashi and Ando 1984), or from the primary ectoderm of the caudal germ band (Schwangart 1905; Lautenschlager 1932; Ando and Tanaka 1980; Tanaka 1987; Guelin 1994) (supplementary fig. S3, Supplementary Material online). Taken together, these data suggest that an inductive mechanism may operate to specify germ cells in most moths and butterflies. We speculate that the loss of *oskar* from most lepidopteran genomes may have facilitated or necessitated secondary reversion to the hypothesized ancestral inductive mechanism for germ line specification.

Another order with apparent near-total absence of *oskar* homologs is the Hemiptera (true bugs), whose sister group Thysanoptera (thrips) nevertheless possesses *oskar*. This secondary loss of *oskar* from a last common hemipteran ancestor correlates with the reported postblastoderm appearance of primordial germ cells in the embryo. Classical studies on most hemipteran species describe germ cell formation as occurring after cellular blastoderm formation, on the inner (yolk-facing) side of the posterior blastoderm surface (Metschnikoff 1866; Witlaczil 1884; Will 1888; Mellanby 1935; Butt 1949; Kelly and Huebner 1989; Heming and Huebner 1994). A notable exception to this is the parthenogenetic pea aphid *Acyrthosiphon pisum*, for which strong gene expression and morphological evidence supports a germ plasm-driven germ cell specification mechanism in both sexual and asexual modes (Miura et al. 2003; Chang et al. 2006; Lin et al. 2014). In contrast, studies of the aphids *Aphis plantoides*, *A. rosea*, and *A. pelargonii* describe no germ plasm, and postblastoderm germ cell formation (Metschnikoff 1866; Witlaczil 1884; Will 1888). However, the genomes of all aphids studied here, including *A. pisum* and three *Aphis* species, appear to lack *oskar*. This suggests that germ plasm assembly in *A. pisum* either does not require a nucleator molecule or uses a novel non-Oskar nucleator.

In the Hymenoptera (ants, bees, wasps, and sawflies), our results strongly suggest that *oskar* was lost from the genome of the last common ancestor of bees and spheroid wasps (supplementary fig. S12, Supplementary Material online). Our analysis further suggests multiple additional independent losses in as many as 25 other hymenopteran lineages, including some for which good quality RefSeq genomes were available (e.g., the slender twig ant *Pseudomyrmex gracilis* or the wheat stem sawfly *Cephus cinctus*) (supplementary fig. S12, Supplementary Material online). However, it would be premature to draw strong conclusions about the number of independent losses given the predominance of transcriptome data in the Hymenoptera.

In addition to convergent losses of *oskar*, we also found evidence for clade-specific duplications of *oskar* in the Hymenoptera. Seven of the nine families containing these putative duplications are families of parasitoid wasps; the remaining two families are ants (Formicidae) and the group of yellowjackets, hornets, and paper wasps (Vespidae) (fig. 4). The phylogenetic relationships of these groups make it highly unlikely that a duplication occurred only once in their last common ancestor, which would be the last common

ancestor of all wasps, bees, and ants (i.e., Apocrita, all hymenopterans except sawflies) (supplementary fig. S12, Supplementary Material online). We suggest that the most parsimonious hypothesis is one of three to five independent duplications of *oskar*, followed by at least 9−14 independent reversions to a single copy, or total loss of the locus (supplementary fig. S9, Supplementary Material online).

No notable life history characteristics appear to unite those species with multiple *oskar* homologs: They include eusocial and solitary, sting-bearing and stingless, parasitoid and nonparasitic insects. To our knowledge, neither is there anything unique about the germ line specification process in Hymenoptera with one or more than one *oskar* homolog. Most Hymenoptera appear to use a germ plasm-driven mechanism to specify germ cells in early blastoderm stage embryos (supplementary fig. S12 and references therein, Supplementary Material online), and we identified *oskar* homologs for all such species described in the embryological literature (supplementary fig. S12, Supplementary Material online). In the notable example of the honeybee *A. mellifera*, in which cytological and molecular evidence suggests germ cell arise from abdominal mesoderm (Bütschli 1870; Nelson 1915; Fleig and Sander 1985, 1986; Zissler 1992; Gutzeit et al. 1993; Dearden 2006), we identified no *oskar* homolog in its well-annotated genome (supplementary fig. S12, Supplementary Material online), as noted previously by other authors (Lynch et al. 2011). However, no major differences in germ plasm or pole cell formation have been reported in species or families of ants or wasps with duplicated *oskar* loci, compared with close relatives that possess *oskar* in single copy [e.g., compare the ants *Solenopsis invicta* (at least two *oskars*) and *Aphaenogaster rudis* (one *oskar*) (Khila and Abouheif 2008), or the pteromalid wasps *Nasonia vitripennis* (one *oskar*) (Lynch and Desplan 2010; Lynch et al. 2011; Quan et al. 2019) and *Otitesella tsamvi* (two *oskars*)]. Thus, future studies that independently abrogate the functions of each paralog individually, will be needed to determine the biological significance, if any, of these *oskar* duplications.

Evolution of the Long Oskar Domain

We showed that the Long Oskar domain is an evolutionary novelty confined to a subset of Diptera. This raises the question of whether the evolution of this domain led to any novel functional properties of *oskar* in these Diptera, relative to its functions in other insects. The only data available on the specific functions of the Long Oskar domain are from studies on *D. melanogaster*. The Long Oskar (606 amino acids: possessing the Long Oskar domain) and Short Oskar (468 amino acids: lacking the Long Oskar domain) isoforms are generated by translation of *oskar* mRNA from alternate initiation codons within the same transcript (Markussen et al. 1995). Short Oskar alone cannot maintain *oskar* mRNA or either protein isoform at the posterior pole of the oocyte or embryo (Vanzo and Ephrussi 2002). However, Short Oskar alone is able to promote the formation of pole cells, albeit many fewer than wild type (Markussen et al. 1995). In contrast, Long Oskar alone can anchor *oskar* mRNA, Oskar protein, and mitochondria at

the posterior pole, but cannot promote pole cell formation (Rongo et al. 1997; Vanzo and Ephrussi 2002; Hurd et al. 2016). In vitro, Short Oskar has a higher affinity for germ plasm components than Long Oskar (Breitwieser et al. 1996; Babu et al. 2004; Anne and Mechler 2005; Megosh et al. 2006; Suyama et al. 2009; Anne 2010). Furthermore, Short Oskar associates with the cytoplasmic germ granules themselves, whereas Long Oskar instead associates with endosomal membranes (Vanzo et al. 2007). These observations have led to the model that Long Oskar's main role is to recruit and anchor Short Oskar to the posterior, where Short Oskar is responsible for germ plasm assembly per se (Markussen et al. 1995; Vanzo and Ephrussi 2002; Tanaka and Nakamura 2008; Tanaka et al. 2011; Hurd et al. 2016).

The molecular basis for the apparently distinct roles of these two isoforms remains largely unclear, and is unlikely to reside entirely within the Long Oskar domain. In vivo assessments of the 139-amino acid Long Oskar domain alone show that it is necessary and sufficient to maintain mitochondria at the oocyte cortex (Hurd et al. 2016). This Long Oskar domain-mediated mitochondrial maintenance requires an intact F-actin cortical cytoskeleton, which is modified by the presence of the Long Oskar domain (Tanaka and Nakamura 2008; Tanaka et al. 2011; Hurd et al. 2016). Compared with controls, long oskar null mutant flies (possessing only Short Oskar) generate fewer PGCs with fewer mitochondria, and their ovaries lack germ cells more often than controls (Hurd et al. 2016).

Although the Long Oskar isoform thus appears to play important and unique roles in functional germ plasm assembly in *D. melanogaster*, these roles appear to be performed perfectly well by the single isoform possessed by nearly all other insects, which in terms of sequence is essentially equivalent to Short Oskar. One or more of posterior *oskar* and germ plasm localization, posterior pole cell formation, and mitochondrial enrichment within germ plasm have been reported for species of ants, bees, wasps, beetles, mosquitoes, and flies that all lack a Long Oskar isoform (Nardon 1971; Jaglarz et al. 2003; Goltsev et al. 2004; Zhurov et al. 2004; Juhn and James 2006; Nardon 2006; Juhn et al. 2008; Lynch et al. 2011; Yoon et al. 2019; Rafiqi et al. 2020). We note, however, that many of these species are reported to possess an oosome, which is a single, morphologically distinct discrete nonmembrane-bound organelle that houses germ plasm components (Meng 1968; Nardon 1971; Klag and Bilinski 1993; Jaglarz et al. 2003; Zhurov et al. 2004; Nardon 2006; Lynch et al. 2011; Quan et al. 2019). This is distinct from most *Drosophila* species for which data are available, whose germ plasm is in the form of multiple smaller granules loosely clustered near the posterior cortex (Mahowald 1962, 1968; Mahowald et al. 1976). We therefore speculate that the evolution of the Long Oskar domain may have enabled tight cortical anchoring of germ plasm components via interaction with endosomes and/or the F-actin cytoskeleton, eliminating the need for an oosome to ensure integrity or local concentration of germ plasm.

Reexamination of Potential Interactions between the LOTUS Domain and RNA

Proteins with a LOTUS domain commonly participate in nucleic acid binding (Williams et al. 1993; Gajiwala and Burley 2000; Liu et al. 2001; Aravind et al. 2005; Lachke et al. 2011; Cui et al. 2013; Harami et al. 2013; Mukherjee et al. 2014). LOTUS domain-containing proteins, particularly RNA-binding proteins (Cui et al. 2013), are often enriched in germ plasm (Anantharaman et al. 2010; Callebaut and Mornon 2010), as are specific RNAs (Ephrussi et al. 1991; Wang and Lehmann 1991; Jongens et al. 1992; Smith et al. 1992; Kobayashi et al. 1995; Nakamura et al. 1996; Mahowald 2001; Vanzo and Ephrussi 2002; Ewen-Campen et al. 2010). However, to date there is no direct evidence that Oskar's LOTUS domain interacts directly with RNA. We were therefore intrigued to find that our bioinformatic analysis suggested that the LOTUS helix α5 might have binding RNA ability (fig. 6d). Consistent with the possibility that Oskar's LOTUS domain might somehow interact with RNA in vivo, we have observed that a loss of function *oskar* allele lacking the entire LOTUS domain (*oskar[ΔLOTUS]*), is unable to direct accumulation of Nanos protein in the germ plasm (Extavour lab, unpublished observation). If the OSK domain, which unlike the LOTUS domain, binds *nanos* mRNA in vitro (Jeske et al. 2015; Yang et al. 2015), were sufficient to ensure Nanos protein localization via *nanos* mRNA recruitment, then germ plasm in *oskar[ΔLOTUS]* flies should contain Nanos protein. Our opposite result could indicate that LOTUS plays a role in RNA binding and/or local translation of *nanos* mRNA. In principle, this could be indirect, for example, aided by LOTUS-mediated oligomerization (Jeske et al. 2015; Yang et al. 2015), or it could be via direct LOTUS–RNA contacts that have not yet been detected in biochemical studies. Further, we note that LOTUS–RNA interactions have, to our knowledge, been probed biochemically and genetically only in *D. melanogaster*, which does not rule out the existence of such binding interactions in other insects.

Functional Implications of Differential Conservation of Regions of the LOTUS and OSK Domains

We have identified novel conserved amino acid positions that we hypothesize are important for the Vasa binding properties of the LOTUS domain and the RNA properties binding of the OSK domain (figs. 6 and 7). Our observation of the conservation of the LOTUS domain α2 helix is consistent with its previously reported importance in LOTUS–Vasa binding (Jeske et al. 2017). In the α2 helix, we also observed high conservation of H227 and Q235. The positions of these residues suggest they may contribute to the interaction between Vasa and LOTUS, but they have not, to our knowledge, yet been implicated functionally in vitro or in vivo. We suggest they should therefore be the target of future mutational studies. Moreover, evolution at the interface between two proteins involves amino acids on both sides of the surface. Therefore, further studies looking at potential coevolution between Oskar and Vasa could shed light on whether the

conserved amino acids that we identified in the LOTUS domain interact with similarly conserved Vasa residues, or whether evolutionary variations in Oskar—Vasa interactions may be explained by coevolution of specific residues at their interaction surfaces (Andreani et al. 2020).

We also uncovered an interesting new conservation pattern within the OSK domain. The conserved amino acids were more abundant in the core of the domain than on the surface. This differential conservation might be relevant to the acquisition of a germ plasm nucleator role of *oskar* in the Holometabla (fig. 5). We noted that the basic properties of surface residues previously reported for *D. melanogaster* (Yang et al. 2015) are conserved across insects, which might indicate that the RNA binding properties of OSK observed in *D. melanogaster* (Jeske et al. 2015; Yang et al. 2015) are also conserved throughout holometabolous insects. We speculate that the comparatively low amino acid conservation of the surface residues in Holometabolous OSK domains, which nevertheless display highly conserved basic properties, could have allowed greater flexibility in the coevolution of specific RNA binding partners for the OSK domains of different lineages.

OSK Evolved Differentially between Holometabolous and Hemimetabolous Insects

Finally, we observed a differential conservation of the OSK domain between hemimetabolous and holometabolous insects. Specifically, we found that the OSK sequence was less conserved across the Holometabola than across the Hemimetabola. This observation raises two potential hypotheses regarding the role of the OSK domain in the functional evolution of Oskar. First, perhaps the apparently relaxed purifying selection experienced by OSK in the Holometabola was necessary for the co-option of *oskar* to a germ plasm nucleation role. Second, Oskar might have a function in the hemimetabolous insects that requires strong conservation of OSK. More studies on the roles and biochemical properties of OSK in hemimetabolous insects will be required to test these hypotheses and further our understanding of the biological relevance of this differential conservation.

In conclusion, analysis of the large data set of novel Oskar sequences presented here provides multiple new testable hypotheses concerning the molecular mechanisms and functional evolution of *oskar*, that will inform future studies on the contribution of this unusual gene to the evolution of animal germ cell specification.

Materials and Methods

Lead Contact and Materials Availability

This study did not generate new unique reagents. This study generated new python3 code and supplementary files referred to below, all of which are available at https://github.com/extavourlab/Oskar_Evolution. Requests for further information and requests for resources and reagents should

be directed to and will be fulfilled by Cassandra G. Extavour (extavour@oeb.harvard.edu).

Experimental Model and Subject Details

This study used no cell culture lines. This study used live samples of *D. melanogaster* and *C. maculatus* and ethanol-preserved samples of *A. asperrimus*. The study also used previously generated genomic and transcriptomic data sets. All the information regarding how those data sets were generated can be found on their respective NCBI pages. The list of all the data sets used in this study can be found in the following files: genome_insect_database.csv, transcriptome_insect_database.csv, genome_crustacean_database.csv, and transcriptome_crustacean_database.csv.

Genome and Transcriptome Preprocessing

We collected all available genome and transcriptome data sets from the NCBI repository registered in September 2019 (fig. 2). NCBI maintains two tiers of genomic data: RefSeq, which contains curated and annotated genomes, and GenBank, which contains nonannotated assembled genomic sequences. Transcriptomes are stored in the transcriptome shotgun assembly (TSA) database, with metadata including details on their origin. Among the registered data sets, five genomes were not yet available, and 40 transcriptomes were only available in the NCBI Trace repository. As they did not comply with the TSA database standards, they were excluded from the analysis. To search for *oskar* homologs in data sets retrieved from GenBank, we needed to generate in silico gene model predictions. We used the genome annotation tool Augustus (Stanke et al. 2006), which requires a hidden Markov model (HMM) gene model. To use HMMs producing gene models that would be as accurate as possible for non-annotated genomes, we selected the most closely related species (species with the most recent last common ancestor) that possessed an annotated RefSeq genome. We then used the Augustus training tool to build an HMM gene model for each genome.

We automated this process by creating a series of python scripts that performed the following tasks:

(1) *1.1_insect_database_builder.py*: This script collects the NCBI metadata regarding genomes and transcriptomes. Using the NCBI Entrez API, it collects the most up to date information on RefSeq, GenBank, and TSA to generate two CSV files: *genome_insect_database.csv* and *transcriptome_insect_database.csv*.

(2) *1.2_data_downloader.py*: This is a python wrapper around the *rsync* tool that downloads the sequence data sets present in the tables created by (1). It automatically downloads all the available information into a local folder.

(3) *1.3_run_augustus_training.py*: This is a python wrapper around the Augustus training tool. It uses the metadata gathered using (1) and the sequence information gathered using (2) to build HMM gene models of all RefSeq

data sets. It outputs sbatch scripts that can be run either locally, or on a SLURM-managed cluster. Those scripts will create unique HMM gene models per species.

At the time of this analysis (September 2019), 133 insect genomes were collected from the RefSeq database, 309 genomes from the GenBank database, and 1,123 transcriptomes from the TSA database. All the accession numbers and metadata are available in the two tables (*genome_insect_database.csv* and *transcriptome_insect_database.csv*) provided in the supplementary files. This pipeline was repeated for crustaceans and the information can be found in the following two files: *genome_crustacean_database.csv* and *transcriptome_crustacean_database.csv*.

Creation of Protein Sequence Databases

The classical approach for homology detection compares protein sequences to amino acid HMM corresponding to the gene of interest. Since we used three different NCBI databases, we performed the following preprocessing actions:

(1) RefSeq: Well-annotated genomes from NCBI contain gene model translation; no extra processing was required.

(2) GenBank: Using the HMMs created from the RefSeq databases, we created gene models for each GenBank genome using Augustus and a custom HMM gene model. To choose which HMM gene model to use, we selected the one for each insect order that had the highest training accuracy. In the case where an insect order did not have any member in the RefSeq database, we used the model of the most closely related order. We then translated the inferred coding sequences to create a protein database for each genome. The assignment of the models used to infer the proteins of each GenBank genome is available in the *Table_S4_models.csv*, Supplementary Material online, available through the GitHub repository for this study at https://github.com/extavourlab/Oskar_Evolution. To automate the process, we created a custom python script available in the file 1.4_run_augustus.py.

(3) TSA: Transcriptomes were translated using the emboss tool Transeq (Madeira et al. 2019). We used this tool with the default parameters, except for the six-frame translation, trim and clean flags. This generated amino acid sequences for each transcript and each potential reading frame.

Identification of Oskar Homologs

The *oskar* gene is composed of two conserved domains, LOTUS and OSK, separated by a highly variable interdomain linker sequence (Ahuja and Extavour 2014; Jeske et al. 2015; Yang et al. 2015). To our knowledge, no other gene reported in any domain of life possesses this domain composition (Blondel et al. 2020). Therefore, here we use the same definition of *oskar* homology as in our previous work: a sequence possessing a LOTUS domain followed by an interdomain region, and then an OSK domain (Blondel et al. 2020). To maximize the number of potential homologs, we searched each

sequence with the previously generated HMM for the LOTUS and OSK domains (Blondel et al. 2020). The presence and order of each domain were then verified for each potential hit and only sequences with the previously defined Oskar structure were kept for further processing. We used the HMMER 3.1 tool suite to build the domain HMM (*hmmbuild* with default parameters), and then searched the generated protein databases (see Creation of Protein Sequence Databases) using those models (*hmmsearch* with default parameters). Hits with an E-value $\geq$ 0.05 were discarded. A summary of all searches performed is compiled in *Table_S5_searches.csv* Supplementary Material online, in the GitHub repository for this study at https://github.com/extavourlab/Oskar_Evolution.

All the hits were then aligned with *hmmalign* with default parameters and the HMM of the full-length Oskar alignment previously generated (Blondel et al. 2020). The resulting sequences were automatically processed to remove assembly artifacts, and potential isoforms. This filtration step was automated and went as follows: First, the sequences were grouped by taxon. Then each group of sequences was aligned using MUSCLE (Edgar 2004) with default parameters. The Hamming distance (Hamming 1950), a metric that computes the number of different letters between two strings, between each sequence in the alignment, was computed. If any group of sequences had a Hamming distance of > 80%, then we only kept the sequence with the lowest E-value match. This created a set of sequences containing multiple *oskar* homologs per species only if they were the likely product of a gene duplication event. We then used the resulting new alignment to generate a new domain HMM and a new full-length Oskar HMM (using *hmmbuild* with default parameters) and ran further iterations of this detection pipeline until we could detect no new *oskar* homologs in the available sequence data sets. We called this final set the filtered set of sequences and used it in all subsequent homology analyses unless otherwise specified.

The Oskar sequences obtained are available in the following supplementary files: *Oskar_filtered.aligned.fasta*, *Oskar_filtered.fasta*, and *Oskar_consensus.hmm*.

The domain definitions for the LOTUS and OSK domains are available in the following supplementary files: *Oskar_filtered.aligned.LOTUS_domain.fasta*, *LOTUS_consensus.hmm*, *Oskar_filtered.aligned.OSK_domain.fasta*, *OSK_consensus.hmm* (see 1.5_Oskar_tracker.ipynb).

Correlative Analysis of Assembly Quality and Absence of Oskar

Using the metadata gathered previously from NCBI databases (see Genomes and Transcriptomes Preprocessing) we created two pools of source data: genomes where we identified an *oskar* sequence, and genomes where we failed to find a sequence that met our homology criteria. We then compared the two distributions for each of the eight available assembly statistics: 1) Contig and 2) Scaffold N50, 3) Contig and 4) Scaffold L50, 5) Contig and 6) Scaffold counts, and 7) Number of Contigs and 8) Scaffolds per genome length.

Finally, we performed a Mann−Whitney U statistical analysis to compare the means of the two distributions (see 2.1_Oskar_discovery_quality.ipynb).

TSA Metadata Parsing and Curation

Data sets in the TSA database are associated with a biosample object that contains all the metadata surrounding the RNA sequencing acquisitions. These metadata can include information about one or both the tissue of origin and the organism's developmental stage. We first automated the retrieval of these metadata using a custom python script that used the NCBI Entrez API (see 2.3_Oskar_tissues_stages.ipynb). However, the metadata proved to be complex to parse for the following reasons: 1) not all projects had the data entered in the corresponding tag, 2) some data contained typographical errors, and 3) multiple synonyms were used to describe the same thing with different words in different data sets. We therefore created a custom parsing and cleaning pipeline that corrected mistakes and aggregated them into a cohesive set of unique terms that we thought would be most informative to interpret the presence or absence of oskar homologs (see 2.3_Oskar_tissues_stages.ipynb to see the mapping table). This strategy sacrificed some of the fine-grained information contained in custom metadata (e.g., "right leg" became "leg") but allowed us to analyze the expression of oskar using consistent criteria throughout all the data sets. This pipeline generated, for all available data sets, a table of tissues and developmental stages including oskar presence or absence in the data set (see Oskar_all_tissues_stages.csv).

Dimensionality Reduction of Oskar Alignment Sequence Space

The Oskar alignment was subjected to an MCA. Similar to a PCA, dimension vectors were first computed to maximize the spread of the underlying data in the new dimensions, except that instead of a continuous data set, each variable (here an amino acid at a given position) contributes to the continuous value on that dimension. Once the projection vectors were computed, each sequence was then mapped onto the dimensions. Each amino acid position (column) in the alignment was considered a dimension with a possible value set of 21 (20 amino acids and gap). We first removed the columns of low information (columns that had <30% amino acid occupancy) using trimal (Capella-Gutierrez et al. 2009) with a cutoff parameter set at 0.3. Then, the alignment was decomposed into its eigenvectors, and projected to the first three components. To perform this decomposition, we implemented a previously developed preprocessing method (Rausell et al. 2010) in a python script (see MCA.py and 2.8_Oskar_MCA_Analysis.ipynb) and performed the eigenvector decomposition with the previously developed MCA python library (see Key Resource Table). We ran the same algorithm on the LOTUS domain, OSK domain, and full-length Oskar alignments obtained above (see Identification of oskar Homologs).

Phylogenetic Inference of Oskar Sequences in the Hymenoptera

We aligned all hymenopteran Oskar sequences using PRANK (Loytynoja 2014) with default parameters. We then manually annotated duplicated sequences by considering two sequences from the same species that had < 80% amino acid identity, as within-species duplications of oskar. We trimmed this alignment to remove all columns with < 50% occupancy using trimal with the cutoff parameter set at 0.5. To reconstruct the phylogeny of these sequences, we used the maximum likelihood inference software RAxML (Stamatakis 2014) with a gamma-distributed protein model, and activated the flag for auto model selection. We ran 100 bootstraps and then visualized and annotated the obtained tree with Ete3 (Huerta-Cepas et al. 2016) in a custom ipython notebook (see 2.7_Oskar_duplication.ipynb).

Calculation of Oskar Conservation Scores

Using the large set of homologous Oskar sequences obtained as described above, we computed different conservation scores for each amino acid position. This methodology relies on the hypothesis that if an amino acid, or its associated chemical properties at a particular position in the sequence are important for the structure and/or function of the protein, they will be conserved across evolution. We considered multiple conservation metrics, each highlighting a particular aspect of the protein's properties as described in the following sections. The scores can be found in the supplementary file scores.csv.

Computation of the Valdar Score

The Valdar score (Valdar 2002) attempts to account for transition probabilities, stereochemical properties, amino acid frequency gaps, and, particularly essential for this study, sequence weighting. Due to the heterogeneity of sequence data set availability, most Oskar sequences occupy only a small portion of insect diversity, primarily Hymenoptera and Diptera. Sequence weighting allows for the normalization of the influence of each sequence on the score based on how many similar sequences are present in the alignment (Valdar 2002). We implemented the algorithm described in Valdar (2002) in a python script (see besse_blondel_conservation_scores.py), then calculated the conservation scores for the Oskar alignment we generated above.

Computation of the Jensen–Shannon Divergence Score

Jensen–Shannon Divergence (JSD) (Lin 1991; Capra and Singh 2007) uses the amino acid and stereochemical properties to infer the "amount" of evolutionary pressure an amino acid position may be subject to. This score uses an information theory approach by measuring how much information (in bits) any position in the alignment brings to the overall alignment (Capra and Singh 2007). This score also takes into account neighboring amino acids in calculating the importance of each amino acid. We used the previously published python code to calculate the JSD of our previously

generated Oskar alignment ([Capra and Singh 2007](#)) (see score_conservation.py).

Computation of the Conservation Bias

The measure of differences in conservation between the holometabolous and hemimetabolous Oskar sequences presented in the results was done as follows: We first split the alignment into two groups containing the sequences from each clade (see 2.4_Oskar_pgc_specification.ipynb). Due to the high heterogeneity in taxon sampling between hemimetabolous and holometabolous insects, we ran a bootstrapped approximation of the conservation scores on holometabolous sequences. We randomly selected N sequences (N = the number of hemimetabolous sequences), computed the Valdar conservation score (see Computation of the Valdar Score), and stored it. After 1,000 iterations, we computed the mean conservation score for each position for holometabolous sequences. For hemimetabolous sequences, we directly calculated the Valdar score using the method as described above (see Computation of the Valdar Score). For each position, we then computed what we refer to as the "conservation bias" between Holometabola and Hemimetabola by taking the ratio of the log of the conservation score Holometabola and Hemimetabola. $Conservation\ Bias = \frac{Log(Valdar_{holo})}{Log(Valdar_{hemi})}$ for each position (see 3.4_LogRatio_Bootstrap.ipynb).

Computation of the Electrostatic Conservation Score

To study the conservation of electrostatic properties of the Oskar protein we computed our own implementation of an electrostatic conservation score (see besse_blondel_conservation_scores.py). Aspartic acid and Glutamic acid were given a score of -1, Arginine and Lysine a score of 1, and Histidine a score of 0.5. All other amino acids were given a score of 0. Then, we summed the electrostatic score for each sequence at each position and divided this raw score by the total number of sequences in the alignment. This computation assigns a score between -1 and 1 at each position, -1 being a negative charge conserved across all sequences, and 1 a positive charge.

Computation of the Hydrophobic Conservation Score

To study the conservation of hydrophobic properties of the Oskar protein we implemented our own hydrophobic conservation score (see besse_blondel_conservation_scores.py). At each position, each amino acid was given a hydrophobic score taken from a previously published scoring table ([Moon and Fleming 2011](#)). (This table is implemented in the besse_blondel_conservation_score.py file for simplicity.) Scores at each position were then averaged across all sequences. This metric allowed us to measure the hydrophobicity conservation of each position in the alignment and is bounded between 5.39 and -2.20.

Computation of the RNA Binding Affinity Score

RNA binding sites are defined as areas with positively charged residues and hydrophobic residues. To estimate the conservation of RNA binding sites in oskar homologs, we used RNABindR v2.0 ([Terribilini et al. 2007](#)), an algorithm predicting putative RNA binding sites based on sequence information only. We automated the calculation for each sequence by writing a python script that submitted a request to the RNABindR web service (see RNABindR_run_predictions.py). We then aggregated all results into a scoring matrix, and averaged the score obtained for each position. We call this score the RNABindR score and hypothesize that it reflects the conservation of RNA binding properties of the protein. Importantly, this score was obtained in 2017 for only a subset of 219 proteins used in this study (indicated in the supplementary files at: 03_Oskar_scores_generation/RNABindR_raw_sources). Since then, the RNABindR server has been defunct and we could not repeat those measurements as the source code for this software is unavailable.

Computation of Secondary Structure Conservation

Due to the overall low conservation of the LOTUS domain, we decided to see whether the secondary structure was conserved. To this end, we used the secondary structure prediction algorithm JPred 4 ([Drozdetskiy et al. 2015](#)). Given an amino acid sequence, this tool returns a positional prediction for α-helix, β-sheet or unstructured. We used the JPred4 web servers to compute the predictions and processed them into a secondary structure alignment (see 2.6_Oskar_lotus_osk_structures.ipynb). We then used WebLogo ([Crooks et al. 2004](#)) to visualize the conservation of the secondary structure.

Visualization of Conservation Scores

We used PyMOL ([DeLano 2002](#)) to map the computed conservation scores onto the solved structures of LOTUS and OSK ([Jeske et al. 2015](#), [2017](#)). At the time of writing, no full-length Oskar protein structure had been reported. With the caveat that all visualization was done on the structure of the D. melanogaster protein domains, we created a custom python script that augments PyMOL with automatic display and coloring capacities. This script is available as Oskar_pymol_visualization.py, and contains a manual at the beginning of the file. For the OSK domain, we used the structure PDBID: 5A4A, and for the LOTUS domain, PDBID: 5NT7 ([Jeske et al. 2015](#), [2017](#)). The LOTUS structure we used is in complex with Vasa, and in a dimeric form ([Jeske et al. 2017](#)), allowing for easy interpretation of the different conservation scores. For the OSK structure, we removed the residues 399–401 and 604–606 from the PDB file as those amino acids did not align across all sequences and therefore showed highly biased conservation scores.

Statistical Analysis

All statistical analyses were performed using the scipy stats module (https://www.scipy.org/). Significance thresholds for P values were set at 0.05. Statistical tests and P values are reported in the figure legends. All statistical tests can be found in the ipython notebooks mentioned below.

Software and Libraries

All software and libraries used in this study are published under open source libre licenses and are therefore available to any researcher.

Type	Name	Version	Source
Software	HMMER	3.1.b2	http://hmmer.org/
Software	PyMOL	1.8.x	https://pymol.org
Software	rsync	3.1.2	http://rsync.samba.org/
Software	Python 3	3.7	https://www.python.org/
Software	Mrbayes	3.2.6	http://nbisweden.github.io/MrBayes/
Software	trimal	1.2rev59	http://trimal.cgenomics.org/
Software	transeq	6.6.0.0	http://emboss.sourceforge.net/apps/cvs/emboss/apps/transeq.html
Software	augustus	2.5.5	http://augustus.gobics.de/
Software	JPred4	4.0	http://www.compbio.dundee.ac.uk/jpred/
Software	RNABindR	2.0	http://ailab1.ist.psu.edu/RNABindR/
Software	Inkscape	0.92.3	https://inkscape.org/
Library	jupyter	4.4.0	https://jupyter.org/
Library	ete3	3.3.1	http://etetoolkit.org
Library	pandas	0.25.1	https://pandas.pydata.org/
Library	mca	1.0.3	https://pypi.org/project/mca/
Library	fuzzywuzzy	0.17.0	https://github.com/seatgeek/fuzzywuzzy
Library	BeautifulSoup4	4.6.3	https://pypi.org/project/beautifulsoup4/
Library	biopython	1.74	https://pypi.org/project/biopython/
Library	numpy	1.16.2	https://www.numpy.org/
Library	seaborn	0.9.0	https://seaborn.pydata.org/
Library	matplotlib	3.0.0	https://matplotlib.org/
Library	scipy	1.1.0	https://www.scipy.org/
Library	progressbar	3.38.0	https://github.com/niltonvolpato/python-progressbar

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

Data Availability

The study generated a series of python 3 script and python 3 ipython notebook files that perform the entire analysis. All the results presented in this paper can be reproduced by running the aforementioned python 3 code. The primary data, oskar homologs, Oskar alignments, trees, and conservation statistics as well as the code created and used are available as supplementary information. For ease of access, legibility, and reproducibility, the code and data sets have been deposited in a GitHub repository available at https://github.com/extavourlab/Oskar_Evolution (commit ID 4eaaa5b11352277e43da72b98bbad397663293fe).

References

Ahuja A, Extavour CG. 2014. Patterns of molecular evolution of the germ line specification gene *oskar* suggest that a novel domain may contribute to functional divergence in *Drosophila*. *Dev Genes Evol*. 224(2):65–77.

Anantharaman V, Zhang D, Aravind L. 2010. OST-HTH: a novel predicted RNA-binding domain. *Biol Direct*. 5:13.

Ando H, Tanaka M. 1980. Early embryonic development of the primitive moths, *Enduclyta signifer* Walker and *E. excrescens* Butler (Lepidoptera: Hepialidae). *Int J Insect Morphol Embryol*. 9(1):67–77.

Andreani J, Quignot C, Guerois R. 2020. Structural prediction of protein interactions and docking using conservation and coevolution. *Wiley Interdiscip Rev Comput Mol Sci*. 10:e1470.

Anne J. 2010. Targeting and anchoring Tudor in the pole plasm of the *Drosophila* oocyte. *PLoS One* 5(12):e14362.

Anne J, Mechler BM. 2005. Valois, a component of the nuage and pole plasm, is involved in assembly of these structures, and binds to Tudor and the methyltransferase Capsuleen. *Development* 132(9):2167–2177.

Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. 2005. The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev*. 29(2):231–262.

Babu K, Cai Y, Bahri S, Yang X, Chia W. 2004. Roles of Bifocal, Homer, and F-actin in anchoring Oskar to the posterior cortex of *Drosophila* oocytes. *Genes Dev*. 18(2):138–143.

Blondel L, Jones TEM, Extavour CG. 2020. Bacterial contribution to genesis of the novel germ line determinant *oskar*. *eLife* 9:e45539.

Breitwieser W, Markussen F-H, Horstmann H, Ephrussi A. 1996. Oskar protein interaction with Vasa represents an essential step in polar granule assembly. *Genes Dev*. 10(17):2179–2188.

Bütschli O. 1870. Zur Entwicklungsgeschichte der Biene. *Z Wiss Zool*. 20:519–564.

Butt FH. 1949. Embryology of the Milkweed Bug, *Oncopeltus fasciatus* (Hemiptera). *Cornell Exp Station Memoir*. 283:2–43.

Callebaut I, Mornon J-P. 2010. LOTUS, a new domain associated with small RNA pathways in the germline. *Bioinformatics* 26(9):1140–1144.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.

Capra JA, Singh M. 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics* 23(15):1875–1882.

Carter J-M, Baker SC, Pink R, Carter DRF, Collins A, Tomlin J, Gibbs M, Breuker CJ. 2013. Unscrambling butterfly oogenesis. *BioMedCentral Genomics* 14:283–283.

Carter JM, Gibbs M, Breuker CJ. 2015. Divergent RNA localisation patterns of maternal genes regulating embryonic patterning in the butterfly *Pararge aegeria*. *PLoS One* 10(12):e0144471.

Chang CC, Lee WC, Cook CE, Lin GW, Chang T. 2006. Germ-plasm specification and germline development in the parthenogenetic pea aphid *Acyrthosiphon pisum*: vasa and Nanos as markers. *Int J Dev Biol*. 50(4):413–421.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450:203–218.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. 14(6):1188–1190.

Cui G, Botuyan MV, Mer G. 2013. (1)H, (15)N and (13)C resonance assignments for the three LOTUS RNA binding domains of Tudor domain-containing protein TDRD7. *Biomol NMR Assign.* 7(1):79–83.

Dearden PK. 2006. Germ cell development in the honeybee (*Apis mellifera*); *vasa* and *nanos* expression. *BMC Dev Biol.* 6:6.

Dearden PK, Wilson MJ, Sablan L, Osborne PW, Havler M, McNaughton E, Kimura K, Milshina NV, Hasselmann M, Gempe T, et al. 2006. Patterns of conservation and change in honey bee developmental genes. *Genome Res.* 16(11):1376–1384.

DeLano WL. 2002. Pymol: an open-source molecular graphics tool. *CCP4 Newsl Protein Crystallogr.* 40:82–92.

Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* 43(W1):W389–394.

Eastham LES. 1930. The embryology of *Pieris rapae* - Organogeny. *Philos Trans R Soc Lond B Biol Sci.* 219:1–50.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.

Ephrussi A, Dickinson LK, Lehmann R. 1991. Oskar organizes the germ plasm and directs localization of the posterior determinant *nanos*. *Cell* 66(1):37–50.

Ephrussi A, Lehmann R. 1992. Induction of germ cell formation by *oskar*. *Nature* 358(6385):387–392.

Ewen-Campen B, Schwager EE, Extavour CG. 2010. The molecular machinery of germ line specification. *Mol Reprod Dev.* 77(1):3–18.

Ewen-Campen B, Srouji JR, Schwager EE, Extavour CG. 2012. *oskar* predates the evolution of germ plasm in insects. *Curr Biol.* 22(23):2278–2283.

Extavour CG, Akam ME. 2003. Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. *Development* 130(24):5869–5884.

Fleig R, Sander K. 1985. Blastoderm development in honey bee embryogenesis as seen in the scanning electron microscope. *Int J Invertebr Reprod Dev.* 8(4 − 5):279–286.

Fleig R, Sander K. 1986. Embryogenesis of the Honeybee *Apis mellifera* L (Hymenoptera, Apidae) - an SEM Study. *Int J Insect Morphol Embryol.* 15(5 − 6):449–462.

Gajiwala KS, Burley SK. 2000. Winged helix proteins. *Curr Opin Struct Biol.* 10(1):110–116.

Goltsev Y, Hsiong W, Lanzaro G, Levine M. 2004. Different combinations of gap repressors for common stripes in *Anopheles* and *Drosophila* embryos. *Dev Biol.* 275(2):435–446.

Guelin M. 1994. [Activity of W-sex heterochromatin and accumulation of the Nuage in nurse cells of the lepidopteran *Ephestia*]. *C R Acad Sci Paris Ser III.* 317:54–61.

Gutzeit HO, Zissler D, Fleig R. 1993. Oogenesis in the Honeybee *Apis mellifera* - cytological observations on the formation and differentiation of previtellogenic ovarian follicles. *Rouxs Arch Dev Biol.* 202(3):181–191.

Hamming RW. 1950. Error detecting and error correcting codes. *Bell Syst Tech J.* 29(2):147–160.

Harami GM, Gyimesi M, Kovacs M. 2013. From keys to bulldozers: expanding roles for winged helix domains in nucleic-acid-binding proteins. *Trends Biochem Sci.* 38(7):364–371.

Hay B, Jan LY, Jan YN. 1990. Localization of *vasa*, a component of *Drosophila* polar granules, in maternal-effect mutants that alter embryonic anteroposterior polarity. *Development* 109(2):425–433.

Heming BS, Huebner E. 1994. Development of the germ cells and reproductive Primordia in male and female embryos of *Rhodnius prolixus* Stal (Hemiptera, Reduviidae). *Can J Zool.* 72(6):1100–1119.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol.* 33(6):1635–1638.

Hurd TR, Herrmann B, Sauerwald J, Sanny J, Grosch M, Lehmann R. 2016. Long oskar controls mitochondrial inheritance in *Drosophila melanogaster*. *Dev Cell.* 39(5):560–571.

Jaglarz MK, Nowak Z, Biliński SM. 2003. The Balbiani body and generation of early asymmetry in the oocyte of a tiger beetle. *Differentiation* 71(2):142–151.

Jeske M, Bordi M, Glatt S, Muller S, Rybin V, Muller CW, Ephrussi A. 2015. The crystal structure of the *Drosophila* germline inducer oskar identifies two domains with distinct vasa helicase- and RNA-binding activities. *Cell Rep.* 12(4):587–598.

Jeske M, Muller CW, Ephrussi A. 2017. The LOTUS domain is a conserved DEAD-box RNA helicase regulator essential for the recruitment of Vasa to the germ plasm and nuage. *Genes Dev.* 31(9):939–952.

Johannsen OA. 1929. Some phases in the embryonic development of *Diacrisia virginica* Fabr. (Lepidoptera). *J Morphol.* 48(2):493–541.

Jones JR, Macdonald PM. 2007. Oskar controls morphology of polar granules and nuclear bodies in *Drosophila*. *Development* 134(2):233–236.

Jongens TA, Hay B, Jan LY, Jan YN. 1992. The *germ cell-less* gene product: a posteriorly localized component necessary for germ cell development in *Drosophila*. *Cell* 70(4):569–584.

Juhn J, James AA. 2006. *oskar* gene expression in the vector mosquitoes, *Anopheles gambiae* and *Aedes aegypti*. *Insect Mol Biol.* 15(3):363–372.

Juhn J, Marinotti O, Calvo E, James AA. 2008. Gene structure and expression of *nanos* (*nos*) and *oskar* (*osk*) orthologues of the vector mosquito, *Culex quinquefasciatus*. *Insect Mol Biol.* 17(5):545–552.

Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, Gimnich F, Frandsen PB, Zwick A, Dos Reis M, et al. 2019. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc Natl Acad Sci U S A.* 116(45):22657–22663.

Kelly GM, Huebner E. 1989. Embryonic development of the hemipteran insect *Rhodnius prolixus*. *J Morphol.* 199(2):175–196.

Khila A, Abouheif E. 2008. Reproductive constraint is a developmental mechanism that maintains social harmony in advanced ant societies. *Proc Natl Acad Sci U S A.* 105(46):17884–17889.

Kim-Ha J, Smith JL, Macdonald PM. 1991. *oskar* mRNA is localized to the posterior pole of the *Drosophila* oocyte. *Cell* 66(1):23–35.

Kirk DL. 2005. A twelve-step program for evolving multicellularity and a division of labor. *Bioessays* 27(3):299–310.

Klag J, Bilinski S. 1993. Oosome formation in 2 ichneumonid wasps. *Tissue Cell* 25(1):121–128.

Kobayashi S, Amikura R, Nakamura A, Saito H, Okada M. 1995. Mislocalization of *oskar* product in the anterior pole results in ectopic localization of mitochondrial large ribosomal RNA in *Drosophila* embryos. *Dev Biol.* 169(1):384–386.

Kobayashi Y, Ando H. 1984. Mesodermal organogenesis in the embryo of the primitive moth, *Neomicropteryx nipponensis* Issiki (Lepidoptera, Micropterygidae). *J Morphol.* 181(1):29–47.

Lachke SA, Alkuraya FS, Kneeland SC, Ohn T, Aboukhalil A, Howell GR, Saadi I, Cavallesco R, Yue Y, Tsai AC, et al. 2011. Mutations in the RNA granule component TDRD7 cause cataract and glaucoma. *Science* 331(6024):1571–1576.

Lasko P. 2013. The DEAD-box helicase Vasa: evidence for a multiplicity of functions in RNA processes and developmental biology. *Biochim Biophys Acta.* 1829(8):810–816.

Lautenschlager F. 1932. Die Embryonalentwicklung der weiblichen Keimdruse bei der Psychide *Solenobia triquetella*. *Zool Jarh.* 56:121–162.

Lebart L, Morineau A, Warwick KM. 1984. Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Chichester (United Kingdom): John Wiley & Sons.

Lehmann R. 2016. Germ plasm biogenesis–an oskar-centric perspective. *Curr Top Dev Biol.* 116:679–707.

Lehmann R, Nüsslein-Volhard C. 1986. Abdominal segmentation, pole cell formation, and embryonic polarity require the localized activity of *oskar*, a maternal gene in *Drosophila*. *Cell* 47(1):141–152.

Lin GW, Cook CE, Miura T, Chang CC. 2014. Posterior localization of ApVas1 positions the preformed germ plasm in the sexual oviparous pea aphid *Acyrthosiphon pisum*. *EvoDevo* 5:18.

Lin J. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans Inform Theory.* 37(1):145–151.

Liu Y, Manna A, Li R, Martin WE, Murphy RC, Cheung AL, Zhang G. 2001. Crystal structure of the SarR protein from Staphylococcus aureus. *Proc Natl Acad Sci U S A.* 98(12):6877–6882.

Loytynoja A. 2014. Phylogeny-aware alignment with PRANK. *Methods Mol Biol.* 1079:155–170.

Lynch JA, Desplan C. 2010. Novel modes of localization and function of *nanos* in the wasp *Nasonia*. *Development* 137(22):3813–3821.

Lynch JA, Özüak O, Khila A, Abouheif E, Desplan C, Roth S. 2011. The phylogenetic origin of *oskar* coincided with the origin of maternally provisioned germ plasm and pole cells at the base of the Holometabola. *PLoS Genet.* 7(4):e1002029.

Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, et al. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47(W1):W636–W641.

Mahowald AP. 2001. Assembly of the *Drosophila* germ plasm. *Int Rev Cytol.* 203:187–213.

Mahowald AP. 1962. Fine structure of pole cells and polar granules in *Drosophila melanogaster*. *J Exp Zool.* 151(3):201–215.

Mahowald AP. 1968. Polar granules of *Drosophila*. II. Ultrastructural changes during early embryogenesis. *J Exp Zool.* 167(2):237–261.

Mahowald AP, Illmensee K, Turner FR. 1976. Interspecific transplantation of polar plasm between *Drosophila* embryos. *J Cell Biol.* 70(2 pt 1):358–373.

Markussen FH, Michon AM, Breitwieser W, Ephrussi A. 1995. Translational control of *oskar* generates short OSK, the isoform that induces pole plasm assembly. *Development* 121(11):3723–3732.

Matthews BJ, McBride CS, DeGennaro M, Despo O, Vosshall LB. 2016. The neurotranscriptome of the *Aedes aegypti* mosquito. *BioMedCentral Genomics* 17:32.

Megosh HB, Cox DN, Campbell C, Lin H. 2006. The role of PIWI and the miRNA machinery in *Drosophila* germline determination. *Curr Biol.* 16(19):1884–1894.

Mellanby H. 1935. The early embryonic development of *Rhodnius prolixus* (Hemiptera, Heteroptera). *Q J Microsc Sci.* 78:71–90.

Meng C. 1968. Strukturwandel und histochimie Befunde iunbesondere am Oosom wahrend der Oogenese und nach der Ablage des Eies von *Pimpla turionellae* L. (Hymenoptera, Ichenumonidae). *W Roux' Archiv Entwicklungsmechanik.* 161(2):162–208.

Metschnikoff E. 1866. Embryologische Studien an Insekten. *Z Wiss Zool.* 16:389–500.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346(6210):763–767.

Mitter C, Davis DR, Cummings MP. 2017. Phylogeny and evolution of Lepidoptera. *Annu Rev Entomol.* 62:265–283.

Miura T, Braendle C, Shingleton A, Sisk G, Kambhampati S, Stern DL. 2003. A comparison of parthenogenetic and sexual embryogenesis of the pea aphid *Acyrthosiphon pisum* (Hemiptera: Aphidoidea). *J Exp Zool B Mol Dev Evol.* 295(1):59–81.

Miya K. 1953. The presumptive genital region at the blastoderm stage of the silkworm egg. *J Fac Agric Iwate Univ.* 1:223–227.

Miya K. 1958. Studies on the embryonic development of the gonad in the silkworm, *Bombyx mori* L. Part I. Differentiation of germ cells. *J Fac Agric Iwate Univ.* 3:436–467.

Miya K. 1975. Ultrastructural changes of embryonic cells during organogenesis in the silkworm, *Bombyx mori*. I. The Gonad. *J Fac Agric Iwate Univ.* 12:329–338.

Moon CP, Fleming KG. 2011. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proc Natl Acad Sci U S A.* 108(25):10174–10177.

Mukherjee D, Datta AB, Chakrabarti P. 2014. Crystal structure of HlyU, the hemolysin gene transcription activator, from Vibrio cholerae N16961 and functional implications. *Biochim Biophys Acta.* 1844(12):2346–2354.

Nagy L, Riddiford L, Kiguchi K. 1994. Morphogenesis in the early embryo of the Lepidopteran *Bobyx mori*. *Dev Biol.* 165(1):137–151.

Nakamura A, Amikura R, Mukai M, Kobayashi S, Lasko PF. 1996. Requirement for a noncoding RNA in *Drosophila* polar granules for germ cell establishment. *Science* 274(5295):2075–2079.

Nakao H. 1999. Isolation and characterization of a *Bombyx vasa*-like gene. *Dev Genes Evol.* 209(5):312–316.

Nakao H, Hatakeyama M, Lee JM, Shimoda M, Kanda T. 2006. Expression pattern of *Bombyx vasa*-like (BmVLG) protein and its implications in germ cell development. *Dev Genes Evol.* 216(2):94–99.

Nakao H, Matsumoto T, Oba Y, Niimi T, Yaginuma T. 2008. Germ cell specification and early embryonic patterning in *Bombyx mori* as revealed by nanos orthologues. *Evol Dev.* 10(5):546–554.

Nakao H, Takasu Y. 2019. Complexities in *Bombyx* germ cell formation process revealed by Bm-nosO (a *Bombyx* homolog of nanos) knock-out. *Dev Biol.* 445(1):29–36.

Nardon P. 1971. Contribution a l'étude des symbiotes ovariens de *Sitophilus sasakii*: localisation, histochimie et ultrastructure chez la femelle adulte. *C R Acad Sci Ser III Sci Vie.* 272D:2975–2978.

Nardon P. 2006. Ovogenese et transmission des bactéries symbiotiques chez le charancon *Sitophilus oryzae* L. (Coleoptera: Curculionoidea). *Ann Soc Entomol Fr.* 42(2):129–164.

Nelson JA. 1915. The embryology of the honey bee. Princeton (NJ): Princeton University Press.

Noce T, Okamoto-Ito S, Tsunekawa N. 2001. *Vasa* homolog genes in mammalian germ cell development. *Cell Struct Funct.* 26(3):131–136.

Pal D, Chakrabarti P. 2001. Non-hydrogen bond interactions involving the methionine sulfur atom. *J Biomol Struct Dyn.* 19(1):115–128.

Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, Kozlov A, Podsiadlowski L, Petersen M, Lanfear R, et al. 2017. Evolutionary history of the Hymenoptera. *Curr Biol.* 27(7):1013–1018.

Presser BD, Rutschky CW. 1957. The embryonic development of the corn earworm, *Heliothis zea* (Boddie) (Lepidoptea, Phalaenidae). *Ann Entomol Soc Am.* 50(2):133–164.

Quan H, Arsala D, Lynch JA. 2019. Transcriptomic and functional analysis of the oosome, a unique form of germ plasm in the wasp *Nasonia vitripennis*. *BMC Biol.* 17(1):78.

Quan H, Lynch JA. 2016. The evolution of insect germline specification strategies. *Curr Opin Insect Sci.* 13:99–105.

Rafiqi AM, Rajakumar A, Abouheif E. 2020. Origin and elaboration of a major evolutionary transition in individuality. *Nature* 585(7824):239–244.

Rausell A, Juan D, Pazos F, Valencia A. 2010. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci U S A.* 107(5):1995–2000.

Raz E. 2000. The function and regulation of *vasa*-like genes in germ-cell development. *Genome Biol.* 1(3):REVIEWS1017–6.

Rongo C, Broihier HT, Moore L, Van Doren M, Forbes A, Lehmann R. 1997. Germ plasm assembly and germ cell migration in *Drosophila*. *Cold Spring Harb Symp Quant Biol.* LXII:1–11.

Saito. 1937. On the development of the Tusser, *Antheraea pernyi* Guerin-Meneville, with special reference to the comparative embryology of insects. *J Fac Agric Hokkaido Imperial Univ.* 40:35–109.

Schroder R. 2006. *vasa* mRNA accumulates at the posterior pole during blastoderm formation in the flour beetle *Tribolium castaneum*. *Dev Genes Evol.* 216:277–283.

Schwangart F. 1905. Zur Entwickslungsgeschichte der Lepidopteren. *Biol Centralbl.* 25:777–789.

Sehl A. 1931. Furchung und Bildung der Keimanlage bei der Mehlmotte *Ephestia kuehniella*. *Zell Zeit Morph U Okol.* 1:429–506.

Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. *J R Soc Interface.* 11(100):20140419.

Sikosek T, Chan HS, Bornberg-Bauer E. 2012. Escape from adaptive conflict follows from weak functional trade-offs and mutational robustness. *Proc Natl Acad Sci U S A.* 109(37):14888–14893.

Smith JL, Wilson JE, Macdonald PM. 1992. Overexpression of *oskar* directs ectopic activation of *nanos* and presumptive pole cell formation in *Drosophila* embryos. *Cell* 70(5):849–859.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34(Web Server issue):W435–439.

Suyama R, Jenny A, Curado S, Pellis-van Berkel W, Ephrussi A. 2009. The actin-binding protein Lasp promotes Oskar accumulation at the posterior pole of the *Drosophila* embryo. *Development* 136(1):95–105.

Tanaka M. 1987. Differentiation and behaviour of primordial germ cells during the early embryonic development of *Parnassius glacialis* Butler, *Luehdorfia japonica* Leech and *Byasa (Atrophaneura) alcinous* Klug (Lepidoptera: Papilionidae). In: Ando H, Jura C, editors. Recent advances in insect embryology in Japan and Poland. Tsukuba (Japan): Arthropod. Embryol. Soc. Jpn. ISEBU Co. Ltd. p. 255–266.

Tanaka T, Kato Y, Matsuda K, Hanyu-Nakamura K, Nakamura A. 2011. *Drosophila* Mon2 couples Oskar-induced endocytosis with actin remodeling for cortical anchorage of the germ plasm. *Development* 138(12):2523–2532.

Tanaka T, Nakamura A. 2008. The endocytic pathway acts downstream of Oskar in *Drosophila* germ plasm assembly. *Development* 135(6):1107–1117.

Terribilini M, Sander JD, Lee JH, Zaback P, Jernigan RL, Honavar V, Dobbs D. 2007. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.* 35(Web Server issue):W578–584.

Tomaya K. 1902. On the embryology of the silkworm. *Bull College Agric Tokyo.* 5:73–111.

Toshiki T, Chantal CR, Toshio K, Eappen A, Mari K, Natuo K, Jean-Luc T, Bernard M, Gérard C, Paul S, et al. 2000. Germline transformation of the silkworm *Bombyx mori* L. using a piggyBac transposon-derived vector. *Nat Biotechnol.* 18(1):81–84.

Valdar WS. 2002. Scoring residue conservation. *Proteins* 48(2):227–241.

Vanzo N, Oprins A, Xanthakis D, Ephrussi A, Rabouille C. 2007. Stimulation of endocytosis and actin dynamics by Oskar polarizes the *Drosophila* oocyte. *Dev Cell.* 12(4):543–555.

Vanzo NF, Ephrussi A. 2002. Oskar anchoring restricts pole plasm formation to the posterior of the *Drosophila* oocyte. *Development* 129(15):3705–3714.

Wang C, Lehmann R. 1991. Nanos is the localized posterior determinant in *Drosophila. Cell* 66(4):637–647.

Webster PJ, Suen J, Macdonald PM. 1994. *Drosophila virilis oskar* transgenes direct body patterning but not pole cell formation or maintenance of mRNA localization in *D. melanogaster. Development* 120(7):2027–2037.

Whittle CA, Kulkarni A, Chung N, Extavour CG. 2021. Adaptation of codon and amino acid use for translational functions in highly expressed cricket genes. *BMC Genomics* 22(1):234.

Whittle CA, Kulkarni A, Extavour CG. 2021. Evolutionary dynamics of sex-biased genes expressed in cricket brains and gonads. *J Evol Biol.* 34(8):1188–1211.

Will L. 1888. Entwicklungsgescshichte der viviparen Aphiden. *Zool Jarh.* 3:201–280.

Williams SG, Attridge SR, Manning PA. 1993. The transcriptional activator HlyU of Vibrio cholerae: nucleotide sequence and role in virulence gene expression. *Mol Microbiol.* 9(4):751–760.

Witlaczil E. 1884. Entwicklungsgeschichte der Aphiden. *Z Wiss Zool.* 40:559–690.

Woodworth CW. 1889. Studies on the embryological development of *Euvanessa antiopa.* In: Scudder, editor. Butterflies of Eastern United States and Canada. Cambridge, UK. p. 102.

Xu X, Brechbiel JL, Gavis ER. 2013. Dynein-dependent transport of *nanos* RNA in *Drosophila* sensory neurons requires Rumpelstiltskin and the germ plasm Organizer Oskar. *J Neurosci.* 33(37):14791–14800.

Yang N, Yu Z, Hu M, Wang M, Lehmann R, Xu RM. 2015. Structure of *Drosophila* Oskar reveals a novel RNA binding protein. *Proc Natl Acad Sci U S A.* 112(37):11541–11546.

Yoon Y, Klomp J, Martin-Martin I, Criscione F, Calvo E, Ribeiro J, Schmidt-Ott U. 2019. Embryo polarity in moth flies and mosquitoes relies on distinct old genes with localized transcript isoforms. *eLife* 8:e46711.

Zhurov V, Terzin T, Grbic M. 2004. Early blastomere determines embryo proliferation and caste fate in a polyembryonic wasp. *Nature* 432(7018):764–769.

Zissler D. 1992. From egg to pole cells: ultrastructural aspects of early cleavage and germ cell determination in insects. *Microsc Res Tech.* 22(1):49–74.